Residue conservation in the prediction of protein-protein interfaces

William Seth Jermy Valdar

A thesis submitted to the University of London
in the Faculty of Science
for the degree of Doctor of Philosophy

August 2001

Department of Biochemistry
and Molecular Biology
University College London
Gower Street
London WC1E 6BT

Abstract

Evolutionary information derived from the large number of available protein sequences and structures could powerfully guide both analysis and prediction of protein-protein interfaces. Three questions are addressed. First, can residue conservation be quantified? Second, are protein-protein interfaces conserved? Third, can the conservation of protein-protein interfaces be useful in their prediction?

To answer the first question, this work reviews 17 methods to quantify positional residue conservation in multiple alignments. It proposes two new measures: one a concrete score, which is then used throughout the remainder of the work, and the other a generalized formula for scoring conservation.

To answer the second question, the conservation of residues at protein-protein interfaces is compared with other residues on the protein surface in six homodimer families. A probabilistic evaluation shows that interface conservation is higher than expected by chance and usually statistically significant at the 5% level or better.

To answer the third question, the utility of conservation in the discrimination of biological from non-biological crystal contacts is assessed. Conservation and size information is calculated for contacts in 53 families of homodimers and 65 families of monomers. Biological contacts are shown to be usually conserved and typically the largest contact in the crystal. Neural networks are then applied to the problem of using size and conservation alone or in combination to predict whether or not a given contact is biologically relevant. The best neural networks combine the two measures and achieve accuracies of over 98%. It is concluded that although size is the most powerful single discriminant, conservation adds important predictive value.

Acknowledgements

I thank my supervisor Janet Thornton for her guidance, encouragement and for working within the parameters of my eccentricities. I also thank those who have directly collaborated with me in my work, namely Thomas Kabir, Irene Nooren and Hannes Ponstingl.

During my doctoral years I have learned much from those around me. In alphabetical order (with nature of contribution in brackets) I thank the following for their teaching, advice and helpful discussions: Richard Chandler (probability lecturer), Richard Jackson (protein structure), Roman Laskowski (everything, including agony uncle), Andrew Martin (mathematics), Richard Mott (statistics), Phil Scordis (alignments), Hugh Shanahan (mathematics) and Adrian Shepherd (neural networks). I thank all members of the Biomolecular Structure and Modelling group for making my time at UCL comfortable and fun. In particular, I thank Andreas Brakoulias, James Bray and Stuart Rison for their camaraderie. For life support and lively lunches, I owe much (including my sanity) to Rachael Grove, Caroline Hadley, Nick Luscombe, Jane Mabey, Christine Mason and Sarah Teichmann.

Lastly, I thank my parents, Stewart and Jean Valdar, for their love and support during my entire PhD. I dedicate my thesis to them and hope they never ask me to explain its contents.

This work was generously supported by a BBSRC Special Studentship.

Declaration and Copyright Notice

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning. The copyright in text of this thesis rests with the Author.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein-protein interactions

Protein-protein interactions are ubiquitous in biology. Transient associations between proteins underpin a broad range of biological processes, which includes hormone-receptor binding, nuclease inhibition, the action of antibody against antigen, signal transduction, correction of misfolding by chaperones and enzyme allostery. Associations that are more permanent are essential for proteins whose stability or function is defined by a multimeric state. Such proteins range from those in grand assemblies, eg, muscle fibres and viral capsids, to those in humbler ones, eg, oligomeric enzymes and oxygen carriers.

The following section (1.1.1) describes five biological systems. Each system relies on protein-protein interactions in one form or another. These examples illustrate the importance of protein-protein interactions and show that such interactions vary considerably, both in kind and function.

1.1.1 The range and importance of protein-protein interactions

1.1.1.1 Inhibition: barnase and barstar

Barnase is an enzyme secreted by *Bacillus amyloliquefaciens*. It helps provide food for the microorganism by degrading extracellular RNA for ingestion. Although an extracellular enzyme, occasional mistargeting and intracellular folding means some barnase ends up in the cytoplasm. In this compartment, the ribonuclease would destroy the bacterium if were not for the action of barstar, barnase's intracellular inhibitor.

Barstar forms a tight and permanent complex with barnase ($K_d = 10^{-14}$M)(Figure 1.1). It forms salt bridges and hydrogen bonds with the enzyme's catalytic residues, blocking the active site completely and preventing further hydrolysis of RNA. During formation of this enzyme-inhibitor complex, barnase remains relatively rigid. Meanwhile barstar undergoes a conformational change, opening itself up to bind the enzyme (Kleanthous & Pommer, 2000, and refs therein).

1.1.1.2 Signalling: G protein-coupled receptors

Guanine-nucleotide-binding proteins (G proteins) can act as signal transducers. Upon binding guanine nucleotides, such as GTP, a G protein changes its conformation. This in turn alters its ability to interact with different proteins and results in the G protein leaving one protein-protein complex to join another, which propagates a signal. G proteins are implicated in a variety of "information" systems within cells, includ-

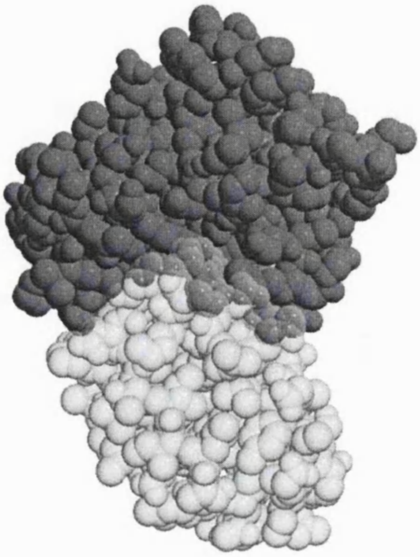


Figure 1.1: Barnase-barstar inhibitor complex.
Barstar (bottom, semitransparent) binds barnase (top, solid dark grey), blocking the enzyme's active site (overlap region). Atom coordinates belong to PDB structure 1brs (Buckle et al, 1994). This picture was generated using MolScript and Raster3D.

ing protein synthesis, cytoskeletal organization, visual transduction and intracellular messenger generation (Morgan, 1994).

G protein-coupled receptors (GPCRs) represent one example of G proteins as signal transducers. In a typical GPCR, such as the β-adrenergic receptor (Figure 1.2), the G protein sits next to the intracellular portion of a hormone receptor. This G protein is heterotrimeric, comprising two tightly associated subunits, G_β and G_γ, and a GTP-binding subunit, G_α. When an agonist hormone, such as adrenaline, binds to the extracellular portion of the GPCR, the intracellular portion changes conformation. This change induces G_α to swap bound GDP for GTP, which in turn causes this subunit to dissociate from the other two. Now "active", G_α may stimulate (or inhibit) a range of downstream targets. Its stimulation of one such target, adenylyl cyclase, promotes the production of cyclic AMP (cAMP) (Hyvönen et al, 2000, and refs therein). The rise in cAMP in turn may stimulate cAMP-dependent kinase and precipitate a cascade of further reactions, which eventually lead to, say, an increase in heart rate.

1.1.1.3 Oligomerization: hydroxylamine oxidoreductase

Nitrification is the bacterial process by which organic nitrogen, in the form of ammonia, is oxidized to nitrite and nitrate. It is part of the biogeochemical nitrogen cycle, which facilitates the exchange of nitrogen between the air, soil and organisms. Hydroxylamine oxidoreductase (HAO) is an important enzyme in nitrification. Once ammonia monooxygenase (AMO) has oxidized ammonia to hydroxylamine (Equation 1.1), HAO oxidizes hydroxylamine to nitrite (Equation 1.2).

$$NH_3 + O_2 + 2H^+ + 2e^- \rightarrow NH_2OH + H_2O \tag{1.1}$$

$$NH_2OH + H_2O \rightarrow HNO_2 + 4H^+ + 4e^- \tag{1.2}$$

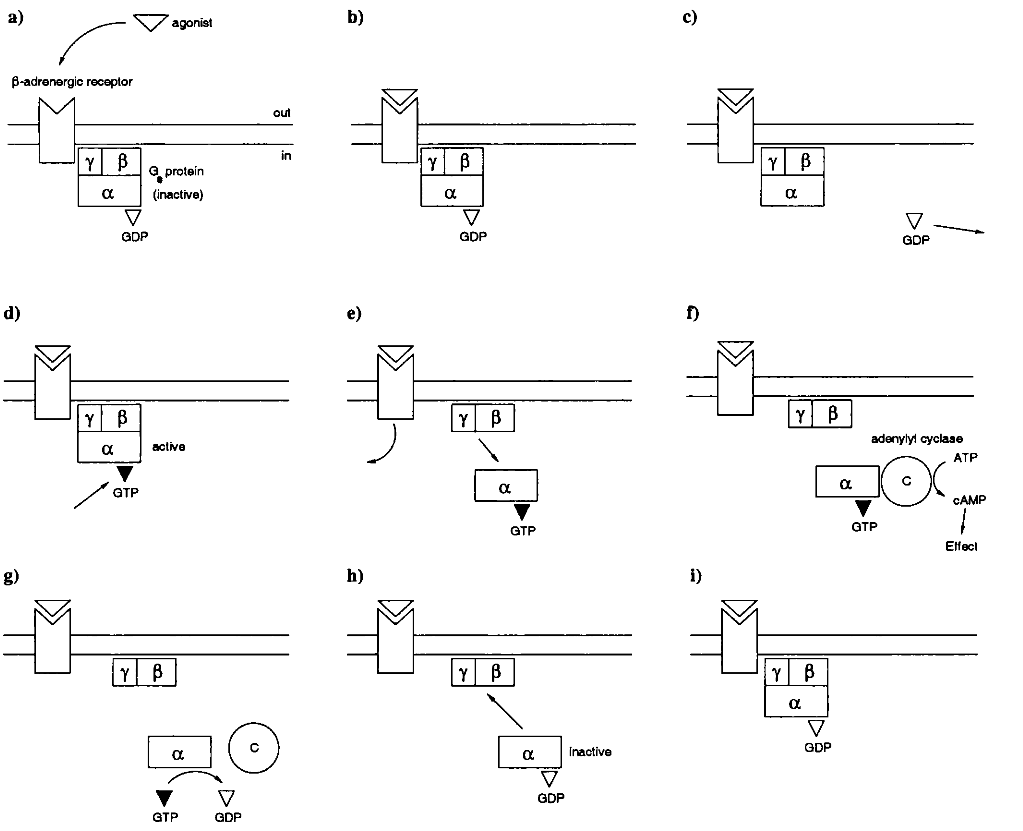


Figure 1.2: Signal transduction in the β-adrenergic G protein-coupled receptor (GPCR). Diagrams (a) to (i) show the order of events when an agonist binds to a β-adrenoceptor that is linked to an adenylyl cyclase-stimulating G protein (G_S). (a) Before agonist-binding, $G_{\alpha s}$ has GDP bound and is inactive. (b) The agonist binds, inducing a conformation change in the GPCR; this causes $G_{\alpha s}$ to exchange GDP (c) for GTP (d) and become active. (e) Now active, $G_{\alpha s}$ has less affinity for the γ and β subunits and dissociates from them. (f) $G_{\alpha s}$ binds allosterically to adenylyl cyclase, promoting cAMP-production. This in turn invokes a chain of further intracellular signals. (g) $G_{\alpha s}$'s low intrinsic GTPase activity means it eventually hydrolyses bound GTP to GDP and returns to an inactive state (h). (i) The now inactive $G_{\alpha s}$ rejoins the γ and β subunits of the GPCR. Adapted from Morgan (Morgan, 1994).

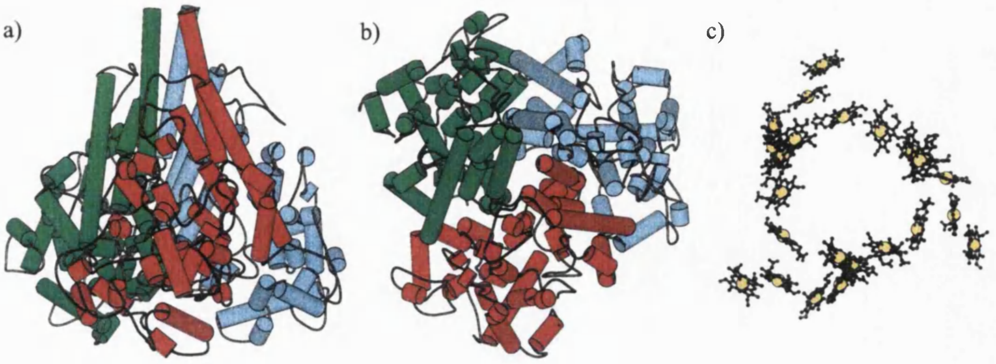


Figure 1.3: Trimeric hydroxylamine oxidoreductase (HAO) from *N. europaea*. The arrangement of peptide chains is shown from the side (a) and top (b). Each subunit is in a different colour: red, green or blue. The 24 haem groups (8 per subunit) are shown separately (c) from the top view. Iron atoms are highlighted in yellow. Trimerization allows the haems to form a ring, which stabilizes electron transfer. Atom coordinates belong to the PDB structure 1fgj (Igarashi et al, 1997) and were obtained from the PQS (Henrick & Thornton, 1998). The figure was created using MOLSCRIPT (Kraulis, 1991).

These two reactions occur in the autotrophic bacterium *Nitrosomonas europaea* and, by donating electrons to its respiratory electron transfer chain, provide energy for this microorganism's growth (Richardson & Watmough, 1999, and refs therein).

HAO is a homotrimer (Figure 1.3). Each subunit contains eight haem groups. Haem groups are useful for redox reactions such as 1.2 because, typically, a single haem can bind and transfer one electron at a time. The haem groups of HAO are not typical. Their sophisticated arrangement supports an electron transfer network that can bind and transfer two electrons simultaneously. This so-called "dielectron transfer" allows HAO to oxidize NH_2OH (reaction 1.3) in a more efficient two steps (reactions 1.3 and 1.4).

$$NH_2OH \rightarrow (HNO) + 2H^+ + 2e^-, \quad (1.3)$$

$$(HNO) + H_2O \rightarrow HNO_2 + 2H^+ + 2e^-. \quad (1.4)$$

Trimerization benefits HAO in a number of ways. First, the extensive interfaces between the subunits provide a stable hydrophobic environment for electron transfer. Second, the association of the subunits, which resembles a head of garlic, creates clefts and cavities believed to bind cytochrome c-554, the recipient of this electron transfer. Third, trimerization allows one haem per subunit, known as P460, to crosslink to another subunit. This positioning of P460 is believed central to HAO's catalysis of the two-step reaction activity (Igarashi et al, 1997) (Hendrich et al, 2001).

1.1.1.4 Structural proteins: tubulin, microtubules and dynamic instability

The cytoskeleton is a network of protein filaments that spatially organizes the cytoplasm of eukaryotic cells. It comprises three main types of filaments: actin filaments, microtubules and intermediate filaments. Microtubules are stiff, tube-like polymers of tubulin. These are highly dynamic structures, alternately growing and shrinking by the gain and loss of tubulin subunits. The cell exploits their dynamic behaviour in a number of processes. Among other things, it uses microtubules to position organelles in the cytoplasm, to align chromosomes on the spindle during mitosis and meiosis, and to change its shape in morphogenesis.

The repeated unit that makes up a microtubule is a heterodimer of homologues α-tubulin and β-tubulin

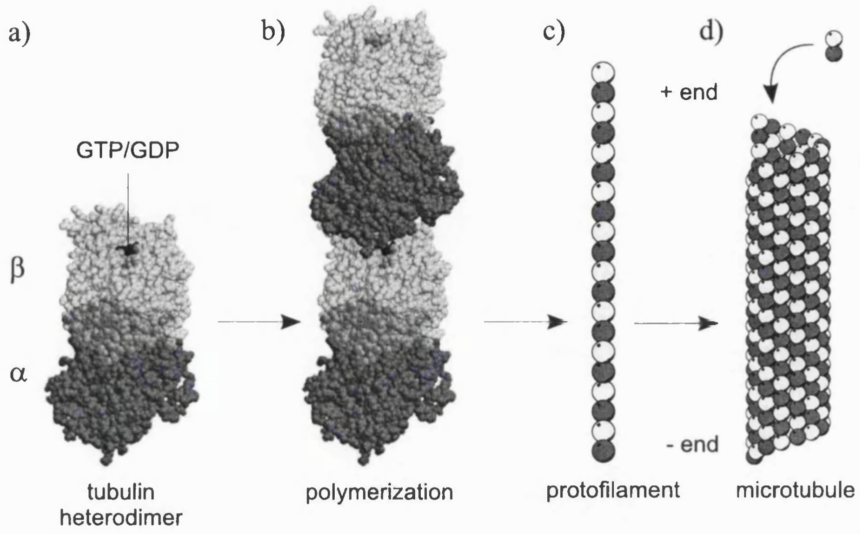


Figure 1.4: Polymerization of tubulin heterodimers into microtubules.
(a) Tubulin heterodimer: α-subunit (dark grey, bottom), β-subunit (semitransparent light grey, top) and site of GTP or GDP binding. (b) Polymerization: the α-subunit of a new unit binds to the β-subunit of an existing unit. (c) Tubulin units polymerize to form protofilaments. (d) Protofilaments assemble into hollow polymers (microtubules), 13 protofilaments in cross-section. (a) and (b) depict PDB structure 1ffx (Gigant et al, 2000). Images were created using MolScript and Raster3D.

(Figure 1.4). These heterodimers associate end-to-end to form protofilaments, which combine laterally to make tubes. Because heterodimers are asymmetric, microtubules are polar: α-tubulin sits at one terminus; β-tubulin sits at the other. This polarity is key to how microtubules grow and how their growth is regulated.

The terminus crowned by β-tubulin is known as the "plus end" of the microtubule because it elongates much faster than the other terminus (the "minus end"). When a new dimer is added to the plus end, its α-subunit interacts with the existing terminal β-subunit. The site of interaction is special for both protomers: it is where the β-subunit normally binds GTP or GDP, and where the α-subunit normally catalyses the hydrolysis of GTP. The rate of polymerization depends on the affinity of the new α-subunit for the existing β-subunit. This in turn depends on whether the β-subunit is currently accommodating a GTP, which acts as bait to the α-subunit, or GDP, which does not. If GTP, the α-subunit binds strongly and microtubule growth is quick. If GDP, the α-subunit binds weakly and no elongation occurs. Meanwhile, α-subunits of dimers already incorporated into the microtubule slowly hydrolyse bound GTP to GDP. This gradually destabilizes the associations between units and can result in depolymerization of the microtubule.

Depending on whether their plus end is capped with GTP or not, microtubules can thus alternate between periods of net growth and net disassembly, a phenomenon called "dynamic instability" (Alberts et al, 1994, and refs therein) (Nogales, 2000, and refs therein).

1.1.1.5 Allostery: haemoglobin and the two-state model

Haemoglobin resides in red blood cells and carries oxygen, as well as CO_2 and H^+, around the bodies of bony vertebrates. Haemoglobin A, the predominant isoform in adults, is a tetramer with the structure $\alpha_2\beta_2$. The α and β subunits are not identical but are homologous and function similarly. Each one contains a single haem prosthetic group, which is responsible for binding O_2. Tetrameric haemoglobin thus carries between zero and four molecules of O_2.

Example interaction	Obligate / non-obligate	Permanent / transient	Specific / multispecific / nonspecific
Barnase-barstar	non-obligate	permanent	specific
G_α-$G_{\beta\gamma}$ complex in GPCR	non-obligate	transient	specific
HAO trimer	obligate	permanent	specific
α-β subunits in tubulin dimer	obligate	permanent	specific
Tubulin units in protofilament	non-obligate	transient	specific
Haemoglobin A tetramer	obligate	permanent	specific
SH3 domain of Abl tyrosine with a range of partners[a]	non-obligate	transient	multispecific
Aggregation of casein	non-obligate	permanent	nonspecific

Table 1.1: Some protein-protein interactions classified.
All interactions shown here are biological.
[a] For example, see Musacchio et al (Musacchio et al, 1994).

Haemoglobin exists in two conformational states: tense (the T-state) and relaxed (the R-state). When haemoglobin is in the R-state, its subunits have a high affinity for O_2. When in the T-state, their affinity is low. Haemoglobin is in continuous equilibrium between these two states, and so the affinity of any one molecule is constantly changing. However, the position of this equilibrium depends on how many O_2 are already bound: the more ligands bound, the further the R-state is favoured and the higher the affinity.

This is an example of "homotropic allostery", ie, when allostery and activity are at equivalent sites in a multimer, and "positive cooperativity", which is illustrated by Equations 1.5 and 1.6:

$$T_n \rightleftharpoons R_n, \quad \text{where } 0 \leq n \leq 4 \tag{1.5}$$

$$K_{\text{equilibrium}} = \frac{[R_n]}{[T_n]} = Lc^n \tag{1.6}$$

where T_n is the T-state with n O_2 ligands bound, c is an affinity factor, and L is the equilibrium constant at T_0 and R_0. The first equation (1.5) shows the equilibrium between the T- and R-states. The second (Equation 1.6) shows how the equilibrium constant for this interconversion is modulated by the value of n. The positive cooperativity of haemoglobin plays an important physiological role, complementing the opposing effects of CO_2 and other heterotropic (ie, binding somewhere other than the active site) allosteric molecules. It helps ensure O_2 tends to be picked up by haemoglobin when it is abundant and tends to be released when it is scarce (Creighton, 1996, and refs therein).

1.1.2 Classifying protein-protein interactions

A particular protein-protein interaction complex may belong to one or more of the following categories: obligate or non-obligate; permanent or transient; biological or crystallographic; specific, multispecific or nonspecific. These classes are defined below and illustrated in Table 1.1.

1.1.2.1 Obligate vs non-obligate

The subunits of an obligate complex are stable and functional within the multimeric state but not outside it. A non-obligate complex is one composed of protomers that are each independently stable and functional in their own right.

1.1.2.2 Permanent vs transient

Transient complexes are those that form temporarily. These are common in signalling pathways and in hormone-receptor binding, and are a type of non-obligate interaction. Permanent complexes can also be non-obligate, but once formed are unlikely to dissociate under normal physiological conditions (eg, enzyme-inhibitor). Obligate complexes are by definition permanent.

1.1.2.3 Biological vs crystallographic

A biological complex is one that exists under normal physiological conditions. A crystallographic complex is one observed in a crystal structure. The interfaces defined in a crystallographic complex are known as "crystal contacts".

Some crystal contacts correspond to real biological interactions. Most do not. Rather, they are artifacts of crystal packing and have no biological relevance. Crystal contacts therefore come in two types:

1. biological crystal contacts, ie, crystal contacts that belong to real biological complexes;
2. nonbiological crystal contacts, ie, those that do not.

In some literature, although not here, "crystal contact" is taken to mean nonbiological crystal contact and "crystallographic complex" is used to describe a crystallographic complex that does not correspond to a biological interaction.

1.1.2.4 Specific, multispecific and nonspecific

The interaction of a protein A with another protein B is specific if both of the following are true:

1. the interaction always occurs at the same site;
2. only B and its analogues bind at this site.

Binding surfaces involved in specific interactions usually exhibit high complementarity. As a result, the complexes formed often have high affinity. All the interactions described in section 1.1.1 are specific. In some literature, although not here, "multispecific" describes a family of homologous proteins that contains members forming different specific interactions. Herein, that version of multispecificity will be termed "familial multispecificity".

The interactions of a protein A with a set of proteins X are multispecific if all of the following are true:

1. the interactions always occur at the same site;
2. only members of X and their analogues bind at this site;
3. X is plural but finite.

Multispecific interactions are not arbitrary; they follow a particular theme. For instance, a typical Src homology 3 (SH3) domain binds a variety of different proteins with low affinity. Nevertheless, it does show some selectivity: it restricts its interactions to targets with a Pro-X-X-Pro binding motif (Mayer, 2001, and refs therein).

The interactions of a protein A are nonspecific if either of the following are true:

1. the interactions occur at random positions A;

2. any protein is a suitable partner.

Aggregation is an example of nonspecific binding: proteins bind to random partners in random orientations.

1.2 Characteristics of interactions and interfaces

Protein-protein interactions occur at the surface of a protein and are biophysical phenomena, governed by the shape, chemical complementarity and flexibility of the molecules involved. Towards the common goal of understanding how proteins interact, a number of studies have characterized the properties of interfaces between polypeptide chains.

A large number of studies of theoretical studies have examined the physical and chemical aspects of protein-protein interfaces. Their findings are described below.

1.2.1 Geometry

1.2.1.1 Size

Most studies define the size of a protein-protein interface with respect to one protomer as the accessible surface area (ASA) lost upon complexation. Estimates of the average size of an interface vary with the type of complex and the dataset used. According to one study by Argos (Argos, 1988), dimers contribute, on average, 12% of their ASA to the contact interface, trimers 17.4% and tetramers 20.9%. These averages conceal great variation. For instance, the interface of dimeric superoxide dismutase is 670 Å ^{2}, 9% of its total ASA; that of tetrameric catalase is 10 570 Å^2, 40% of the surface (Jones & Thornton, 2000, and refs therein). Jones & Thornton (Jones & Thornton, 1996) showed that among homodimers interface ASA is roughly linearly related to molecular weight.

1.2.1.2 Planarity

The planarity of an interface is usually measured as the root-mean-square deviation of its best-fit plane. Most interfaces are relatively flat compared with the rest of the surface. This is particularly the case for homooligomers, partly thanks to packing constraints, and less so for more heterogeneous complexes, eg, antibody-antigen complexes [Thomas Kabir, personal correspondence] (Jones & Thornton, 1997a). There are exceptions. In some obligate complexes, such as the homodimer gamma interferon, subunits interlock in a sprawling, convoluted embrace.

1.2.1.3 Shape

Most interfaces are roughly circular. Jones & Thornton showed that the set of residues in a homodimer interface could be approximated by the set of as many residues in a circular patch. In their dataset, the residue overlap between the theoretical patch and biological interface ranged between 54% and 87% (Jones & Thornton, 1997a).

1.2.1.4 Symmetry

Almost all homooligomeric complexes are symmetrical. There are a number of conjectures as to why this should be. One is that symmetry allows so-called "finite assembly", where the shape and interaction of subunits precludes unwanted aggregation. Blundell & Srinivasan suggest that highly symmetric assemblies are more energetically stable than asymmetric aggregates (Blundell & Srinivasan, 1996). Wolynes has speculated that the folding landscape for symmetric proteins has fewer kinetic barriers (Wolynes, 1996).

Symmetric interfaces may be categorized as isologous or heterologous. Isologous interfaces are those where identical surfaces on two subunits interact. These interfaces have a two-fold axis of symmetry and can occur only in dimers. Heterologous interfaces describe interactions between different surfaces on two subunits. Monod et al have suggested isologous interfaces are easier to evolve from monomers than heterologous ones: a mutation that strengthens the inter-subunit association is acquired doubly if the interface is isologous (Goodsell & Olson, 2000, and refs therein).

1.2.1.5 Complementarity

Geometric complementarity describes the physical fit between two surfaces. Several methods exist to measure geometric complementarity within an interface. Among those used in modern studies are the shape correlation index (Lawrence & Colman, 1993), based on distance and the angle of the normal vectors to the molecular surface, the gap index (Jones & Thornton, 1996), which measures the volume of cavities between the interacting surfaces and normalizes by the interface area, and packing density (Gerstein et al, 1995), measured using Voronoi polyhedra. A study by Lo Conte et al (Lo Conte et al, 1999), which using all three methods, showed that shape complementarity was marginally higher in oligomers and inhibitor complexes than in antibody-antigen complexes, and that packing density at the centre of interfaces resembles that of the protein interior.

1.2.2 Chemistry

1.2.2.1 Hydrophobicity

Hydrophobic interaction is considered a driving force stabilizing associations within proteins (Dill, 1990, and refs therein). However, the hydrophobic effect is thought less influential in associations between proteins (Sheinerman et al, 2000, and refs therein).

In transient complexes, the proportion of hydrophobic groups differs only slightly between the interface and the surface (Lo Conte et al, 1999). In obligate and permanent complexes the interfaces tend to be significantly more hydrophobic (Chothia & Janin, 1975) (Young et al, 1994) (Jones & Thornton, 1995). Studies examining the dispersion of hydrophobicity in interfaces show that, in obligate complexes at least, hydrophobic groups tend to scatter rather than concentrate in a single large patch (Larsen et al, 1997).

1.2.2.2 Amino acid composition

The amino acid composition found at interfaces lies between that of the protein interior and the protein surface but is much closer to the surface. Compared with the rest of the surface, interfaces are typically richer in the aromatic residues His, Tyr, Phe and Trp; somewhat richer in the aliphatic residues Leu, Ile, Val and Met, and depleted in the charged residues Asp, Glu and Lys. Most abundant is the charged residue Arg. This may be because Arg is a prolific hydrogen-bond former and is able to form water-mediated hydrogen bonds with other Args (Magalhaes et al, 1994) (D'Alessio, 1999). Composition varies with the class of interface. Protease-inhibitor complexes are richer in Cys; antibody-antigen interfaces are richer in Tyr. But the preponderance of Arg and depletion of Lys are common (Lo Conte et al, 1999). Figure 1.5 shows the distances between the amino acid compositions of the surface, interior and interface in different classes of complex.

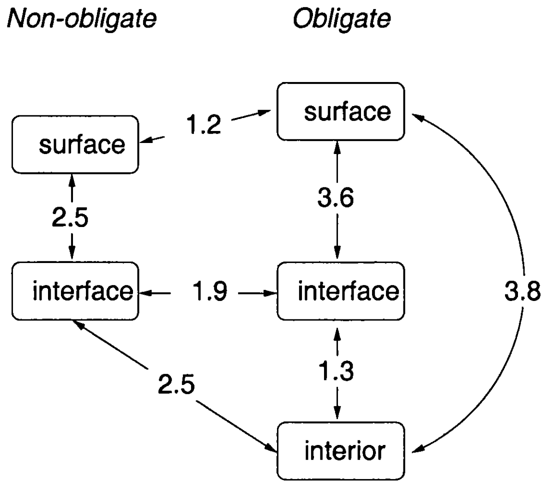


Figure 1.5: Distance between amino acid compositions.
Pairwise distances are calculated by projecting the fractional compositions of amino acids into 20-dimensional space and measuring the Euclidean distance between them, ie, $d = \sqrt{\frac{1}{20}\sum_i (s_i - t_i)^2}$, where s_i and t_i are the percentage areas contributed by residue type i to surfaces s and t . Distance is in units of percentage area (Lo Conte et al, 1999, and Joel Janin, personal correspondence). Note that in this diagram "interface" is defined conservatively and includes only those atoms that are completely buried on complexation. Other studies, such as that of Jones & Thornton (Jones & Thornton, 1995), which tend to define the interface more loosely, show obligate interfaces as being closer in composition to the surface than the interior.

1.2.2.3 Electrostatics

Non-obligate protein-protein interfaces are often quite polar. Polar groups can help make interactions specific: a prospective partner must complement the existing pattern and direction of positive and negative charge. This is an advantage if, as is the case with most non-obligate interactions, subunits must find each other in the cytoplasmic soup. However, polar groups can also oppose complex formation, owing to desolvation effects. A protein-protein interaction must therefore balance the thermodynamic cost of burying polar groups from water with the kinetic benefit of electrostatic interactions at the interface, ie, hydrogen bonds and salt bridges (McCoy et al, 1997) (Sheinerman et al, 2000, and refs therein).

Hydrogen bonds are more common in non-obligate complexes than in obligate or permanent complexes. Average frequencies of hydrogen bonds per 100 Å^2 of buried surface have been calculated as 0.88 for obligate homodimers, 1.4 for permanent non-obligate complexes and 1.1 for antibody-antigen complexes (Jones & Thornton, 2000, and refs therein). However, these averages hide great variation: among non-obligate complexes studying by Lo Conte et al, the number of hydrogen bonds ranged from 3 to 50 (Lo Conte et al, 1999). This study also found the majority of bonds formed were most often mediated by water.

Salt bridges are rarer than hydrogen bonds, occurring in only half the homodimers analysed by Jones & Thornton (Jones & Thornton, 1996). This is unsurprising because ionic interactions, which represent an extreme form electrostatic interaction, can occur between only a subset of hydrogen bonding residues.

Complex	K_d (mol dm^{-3}) range
Ribonuclease inhibitor with angiogenin	10^{-16}
Barnase with barstar	10^{-14}
Caspase-activated DNase (CAD) with inhibitor (ICAD)	10^{-9}
Activated $G_{\alpha s}$ with adenylyl cyclase	10^{-8}
α-tubulin with β-tubulin (in tubulin dimer)	10^{-6}
$\alpha\beta$-haemoglobin with $\alpha\beta$-haemoglobin (in $\alpha_2\beta_2$-haemoglobin)	10^{-6}

Table 1.2: Dissociation constants for some example protein-protein complexes

1.2.3 Prediction of protein-protein interfaces based on physical and chemical characteristics

The consistency apparent in observations of oligomeric interfaces has led some groups to suggest the location of a putative interface may be predictable from protomer structure alone (Young et al, 1994) (Lijnzaad & Argos, 1997) (Jones & Thornton, 1997b). Jones & Thornton (Jones & Thornton, 1997b) developed a predictive method in which, for each protomer in a dataset of dimers, they defined roughly circular patches on the molecular surface, then assessed and ranked each patch according to its chemical and physical properties. Because the properties that make a good interface depend on the type of complex, patches on protomers from homodimer, heterodimer and antigen-antibody complexes were ranked by different criteria. Their method proved most powerful when applied to simple homodimers and weakest when applied to transient dimers, mirroring the degrees of physico-chemical consistency observed for these types of complexes.

The geometric and electrostatic complementarity observed within interfaces has been the basis of many studies that dock two proteins of known structure (Sternberg et al, 1998, and references therein). These algorithms usually begin by treating the two proteins as rigid bodies that are docked to produce a tight complex. Putative complexes are then assessed and refined according to electrostatic or chemical criteria to predict the "best" complex.

1.2.4 Kinetics and energetics of binding

1.2.4.1 Affinity and dissociation constants

The affinity between two protomers in a protein-protein interaction is most often expressed as their dissociation constant at thermodynamic equilibrium, K_d. For the interaction between protomers A and B, ie,

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{AB}\,,$$

K_d is given by

$$K_d = \frac{[\mathrm{A}]\,[\mathrm{B}]}{[\mathrm{AB}]}\,.$$

This scheme also accommodates interactions involving more than two protomers, provided those interactions are first broken down into successive bimolecular steps.

Values of K_d in biological systems range from 10^{-4}, denoting loose association, to 10^{-16}, denoting tight associations. Table 1.2 gives the dissociation constants of some example complexes.

The dissociation constant can be determined experimentally by measuring equilibrium concentrations of A, B and AB. The range of K_d dictates the choice of experimental technique. Available techniques for the micromolar range and above are fluorescence quenching, equilibrium ultracentrifugation or microcalorimetry. For the nanomolar range: enzyme-linked immuno-absorbant assay (ELISA). Below the nanomolar range, direct measurement is unreliable. In this case, kinetic measurements, ie, measurements of association and dissociation rates, are used in preference to equilibrium methods (Janin, 2000, and refs therein).

1.2.4.2 Energetics

The Gibbs free energy of dissociation may be calculated from the K_d using the equation

$$\Delta G_d^{\mathrm{o}} = -RT\ln\frac{K_d}{c^{\mathrm{o}}},$$

where T is temperature, R is the gas constant, and $c^{\mathrm{o}} = 1$ mol dm^{-3} under standard conditions. Although not an energy as such, ΔG_d^{o} is often referred to as the "binding energy". The higher its value binding energy the more favourable the interaction. In biological systems, ΔG_d^{o} ranges from 6 to 19 kcal mol^{-1}.

The binding energy is a balance of two components, entropy and enthalpy. These are related to ΔG_d^{o} by the equation

$$\Delta G_d^{\mathrm{o}} = \Delta H_d^{\mathrm{o}} - T\Delta S_d^{\mathrm{o}},$$

where ΔH_d^{o} and ΔS_d^{o} are the changes in enthalpy and entropy respectively. The interaction is "enthalpy driven" when ΔH_d^{o} is positive (favourable) and ΔS_d^{o} negative (unfavourable). If the converse is true, the interaction is "entropy driven". Determining the relative contributions of enthalpy and entropy to the interaction is not as easy as it might seem. In principle, ΔH_d^{o} could be estimated from van't Hoff's law:

$$\Delta H_d^{\mathrm{o}} = -\mathrm{R}\frac{\mathrm{d}\,(\ln K_d)}{\mathrm{d}\,(1/T)}.$$

However, this works poorly in practice. Rather, ΔH_d^{o} is best measured directly by isothermal titration calorimetry (ITC). In this technique, sensitive microcalorimeters are used to measure the heat evolved on mixing protomers A and B. This heat corresponds to $-\Delta H_d^{\mathrm{o}}$.

The heat capacity change, ΔC_d^{o}, can be calculated by measuring ΔH_d^{o} at different temperatures, thanks to the relationship

$$\Delta C_d^{\mathrm{o}} = \frac{\mathrm{d}\left(\Delta H_d^{\mathrm{o}}\right)}{\mathrm{d}T}.$$

The heat capacity change is useful because it can indicate how much the hydrophobic effect contributes to stabilizing the AB complex. Some groups have postulated a direct relationship between ΔC_d^{o} and the amount of buried hydrophobic surface area of the form

$$\Delta C_d^{\mathrm{o}} = a\Delta\mathrm{ASA}_{nonpolar} + b\Delta\mathrm{ASA}_{polar},$$

where a and b are constants, and $\Delta\mathrm{ASA}_{nonpolar}$ and $\Delta\mathrm{ASA}_{polar}$ are the changes in ASA for nonpolar and polar surfaces respectively. However, there is evidence to suggest the true relationship is less straightforward (Janin, 2000, and refs therein) (Henriques et al, 2000, and refs therein).

1.2.4.3 Anatomy of interface thermodynamics

Some residues at a protein-protein interface contribute more to the binding energy than others. The energetic importance of a residue can be measured by first mutating it to a reference amino acid type, such as Ala, then recording the consequent change in the free energy of binding as

$$\Delta\Delta G_d = \Delta G_d^{\mathrm{wt}} + \Delta G_d^{\mathrm{mut}},$$

where ΔG_d^{wt} and $\Delta G_d^{\mathrm{mut}}$ are values of ΔG_d for the wild-type and mutant respectively. This so-called "alanine-scanning mutagenesis" has been used to map the thermodynamic properties of dimer interfaces and revealed that important residues are not necessarily distributed evenly. Rather, residues with high $\Delta\Delta G_d$ often concentrate in "hot-spots" of binding energy (Bogan & Thorn, 1998). There is also evidence from other thermodynamic studies that residues distant from the interface can play a critical role in stabilizing protein-protein interactions (Hedstrom, 1996). Such residues are believed to be energetically coupled with those directly involved in binding and allow binding energy to propagate through tertiary structure (Lockless & Ranganathan, 1999).

1.3 Experimental technologies for detecting and measuring protein-protein interactions

Protein-protein interactions have been subject to just about every possible form of experimental analysis. Table 1.3 summarizes some of the main experimental methods currently used to study protein-protein interactions.

1.3.1 Classifying experimental methods

Experimental methods for analysing protein-protein interactions are here divided into two classes: presence methods and characterization methods. Presence methods can detect unknown protein-protein interactions. These seek to find out *whether* proteins interact. Characterization methods characterize interactions. They seek to discover *how* proteins interact.

Presence methods provide only general information but can often be performed on a large scale, such as in high-throughput screening. They may be subdivided into three further categories: pairwise methods, which detect whether two proteins interact; fishing experiments, which detect all proteins that interact with a given "bait" protein; and all vs all methods, which detect all interactions among a group of proteins.

Characterization methods provide more detailed information about protein-protein interactions but are usually performed on only a small scale. These are mainly applied to known interactions and illuminate specific aspects of a complex such as binding energy, structure, conformational changes, kinetics and so forth.

1.4 Evolution of oligomers

It is usually assumed that, in evolution, single protomers came first and oligomers later. This can be justified by the intuition that simplicity usually precedes complexity and not the other way around. Some workers have even cited the amino acid composition of oligomeric interfaces as supporting evidence. Specifically, if oligomers came first, then why should oligomeric interfaces be midway in composition between the interior

Experimental technique	Presence[a]			Characterization[a]				Description[b]
	Pairwise	Fishing	All vs all	Binding strength	Hydrophobicity	Kinetics	Structure	
Cell-map proteomics			■					Performed on whole cell or chosen compartments. First isolate protein fraction. Then identify which proteins are present using any of immunoprecipitation, affinity chromatography, sedimentation equilibrium and 1D or 2D gel electrophoresis. Analyse complexes by 1D native gel electrophoresis and mass spectrometry (MALDI-TOF or ES)
Chemical cross-linking			■				■	Covalently link any peptide terminii that are close during complexation
Circular dichroism (CD)							■	Detect changes in secondary structure and thereby show whether conformational changes occur on complexation
Co-crystallography							■	Perform X-ray crystallography on intact complex to give 3D structure of interaction
Electron microscopy (EM)							■	Visualize multimers at a coarse grain level
Isothermal calorimetry (ITC)				■	■			Measure heat change on complexation
Microarrays			■					A microarray is a small chip that contains a matrix of bound cDNA bait. When cell lysate is washed over the chip, mRNA sticks to its complementary bound cDNA. A microarray therefore measures mRNA levels in a cell. Correlated levels of mRNA expression are used as a rough guide as to whether two proteins might interact
Nuclear magnetic resonance (NMR)	■							Isotope labelling: identify close-together atoms in a complex from their NOE transfers
Protein chips			■					Immobilize arrays of protein bait on a chip. Wash solution of (eg, photo-) labelled prey proteins over the chip. Interacting prey protein remain stuck to bait and are detectable by their label
Resonance energy transfer (FRET/BRET)		■				■		Tag subject protein tagged with a fluorescent (FRET) or bioluminescent (BRET) probe. If subject binds to another protein, the probe noticeably changes its wavelength
Surface plasmon resonance (SCR)		■				■		Immobilize bait protein on activated optical metallic surface. Interaction between the bait and prey alters the diffractional properties of the plate, causing an optical change
Yeast two-hybrid (Y2H)			■					In addition to bait and prey, this method involves a reporter gene and two protein domains that bind to it: B, which binds at the promoter, and A, which activates transcription. The bait is fused to B, the prey to A. If the prey and bait interact, so do A and B, which causes expression of the reporter and leads to a colour change or other noticeable phenotype

Table 1.3: Experimental methods for detecting and measuring protein-protein interactions

[a] The binary grid (central columns) classify methods according to their typical use. Black-filled cells indicate membership of a class. See main text for how these classes are defined.

[b] "Bait" and "prey" denote two respective interacting proteins.

and exterior (D'Alessio, 1999, and refs therein)? Accepting this supposition prompts at least two further questions:

1. Does oligomerization confer biological, and therefore evolutionary, advantages?

2. By what mechanisms could oligomeric complexes have arisen from monomers?

This section explores possible answers to the above questions.

1.4.1 Advantages of being an oligomer

Oligomers possess a number of talents not shared by monomers (Goodsell & Olson, 1993, and refs therein) (D'Alessio, 1999, and refs therein). These are summarized below.

1. Subunit interfaces. In some multimeric enzymes (eg, *E. coli* aspartate transcarbamoylase) the active site is formed at the junction of two subunits. This arrangement can bestow several advantages, including improved specificity through substrate channeling and assisted catalysis through subtle intersubunit motions.

2. Interaction of subunits. Subunit interaction can be dynamic. It can endow a complex with cooperativity (eg, in haemoglobin) or substrate feedback inhibition. In the extreme case of bovine mitochondrial F1-ATPase, catalysis is aided by a continuous cycling of the enzyme's oligomeric state (Abrahams et al, 1994).

3. Reduced surface area. Combining several functional proteins into one aggregate reduces the total surface area accessible to solvent. This is easier on the host: fewer ions and ordered water molecules are needed to neutralize and hydrate the protein surface.

4. Multiple active sites. Some multimeric enzymes (eg, dimeric superoxide dismutase) have an active site on each subunit. This reduction of surface area has a kinetic advantage in that productive collisions between enzyme and substrate are more likely.

5. Structural frameworks. Oligomerization enables the cell to form structural frameworks. These can used for protection, scaffolding (eg, tubulin) or mechanical transduction (eg, in muscle contraction).

6. Coding efficiency. Building a range of large structures out of identical bricks, or groups of identical bricks, makes more use of less genetic information.

7. Modular architecture vs one long polypeptide. It is more robust to build a necessarily large and complex structure such as HAO (see 1.1.1.3) from separate modules than from a single polypeptide chain. First, a single chain version of the HAO homotrimer would require a longer gene, which, assuming a constant mutation rate, would accumulate a deleterious mutation more quickly. However, as has been noted by Monod et al (Monod et al, 1965), the effect of a single deleterious mutation in a repeated module would be multiplied throughout the oligomer. Second, if the single chain version of HAO develops a fault during transcription, translation or folding, the whole structure is lost; in the modular scheme, only the faulty module is lost, and this may be replaced.

Points 1, 2, 3 and 4 could also apply to the advantages of having multiple domains in a multidomain protein.

1.4.2 Mechanisms of oligomeric evolution

The most general models of oligomer evolution describe the transition from monomer to dimer (D'Alessio, 1999). This is a reasonable simplification because it can be applied iteratively to explain higher order oligomers. The models described below all assume the following starting condition: the availability of expendable genetic material. This redundancy, which probably entered the hypothetical genome through a duplication event, provides a relatively safe arena for evolutionary experimentation.

1.4.2.1 The mutation model

The mutation model starts with one aloof monomer. This monomer acquires a primary mutation that makes its surface adhesive enough to bind another protein. The site of the mutation need not be the only source of adhesion – it may complement existing "preprimary" mutations – but it tips the balance. The new protein-protein interaction is stable or metastable, which means the complex corresponds to only one of many possible kinetic paths. After this, two things can happen. If the interaction is detrimental to the host organism, it is removed from the gene pool (negative selection). If the interaction is tolerated (neutal survival) or advantageous (positive selection) it stays and continues to evolve. Secondary mutations then make the complex fully stable if it was not so already. Over time, genetic drift leads to the evolved dimer, which is the one observed today. By this point the primary mutation is no longer readily detectable (D'Alessio, 1999, and refs therein).

1.4.2.2 The domain-swap model

The domain-swap model (also impenetrably known as the Rosetta Stone model) starts with two monomers A and B, as shown in Figure 1.6. Fusion of the genes for A and B leads to expression of the fused two-domain protein AB. Relatively few mutations then produce a primitive binding site between the two domains. Successive point mutations optimize this domain-domain interface and result in a stronger association. One of two things may happen next. In the first scenario, recombination separates the genes for A and B, the two domains will once again be separate proteins. However, because of the optimized binding site, monomers A and B will interact as a heterodimer. In the second scenario, a deletion in the loop between the two domains restricts relative positioning of A and B, disrupting the interdomain interface. However, the A and B domains of two AB proteins can interact, forming a domain-swapped homodimer (Marcotte et al, 1999).

The domain-swap model is believed to explain some but not all protein-protein interactions, and considered to be subsumed by the more general mutation model (D'Alessio, 1999).

1.5 The use of sequence data to infer molecular interactions

Sequence-based computational approaches to predicting molecular interactions have become popular in recent years, thanks largely to the rapid growth of available protein and DNA sequence data and related resources. Sequence-based approaches contrast with patch analysis, docking and other computational methods described in section 1.2.3 because they consider the genetic provenance of the interacting proteins or molecules and usually assume some kind of evolutionary model. These methods can be roughly divided into comparative genomic methods, which generally seek to detect potential interacting pairs of proteins, and structure-level methods, which seek to locate or characterize the site of an interaction on a protein structure and the residues involved.

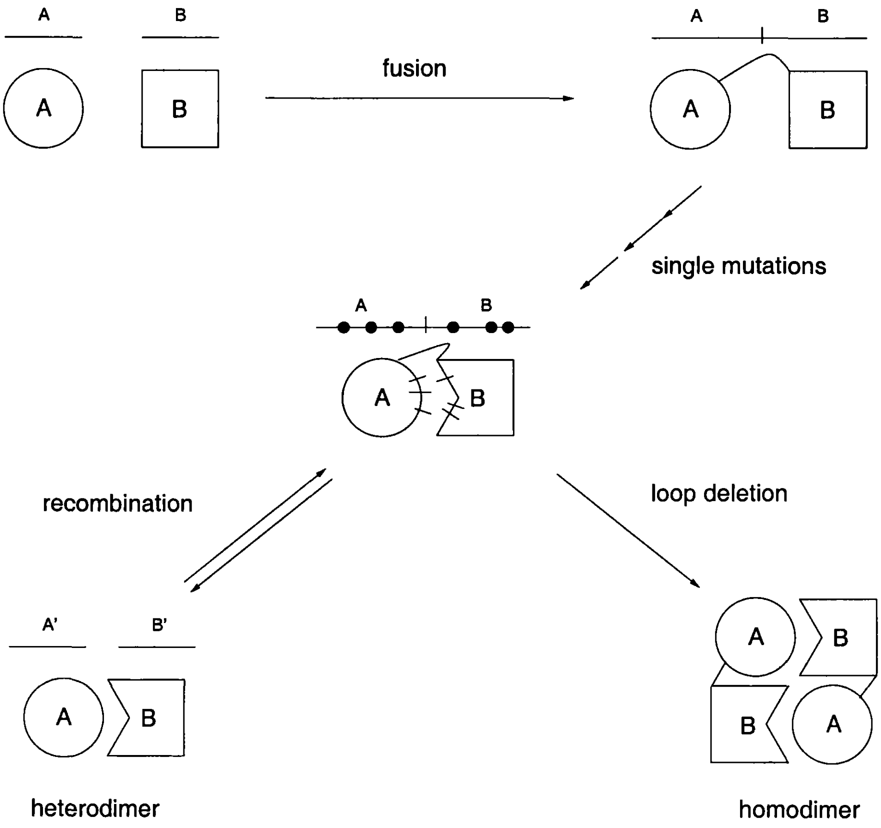


Figure 1.6: domain-swap model of dimer evolution.
Circle A and square B are domains. Lines above shapes represent the genes for A and B. See text for explanation. Adapted from (Marcotte et al, 1999).

1.5.1 Comparative genomics

1.5.1.1 Gene order

The tenet of gene order methods is simple: if two genes are close together, their products are likely to interact. In fact, the closer, the more likely. A study by Huynen et al (Huynen et al, 2000) showed that if the adjacency of two genes was conserved among phylogenetically distant genomes, there is a 63% chance their products are part of the same multimer and a 30% chance they have a direct physical interaction. If one considers only those genes involved in the metabolic pathway of *E. coli*, this likelihood rises to 90%. One explanation for why proteins from adjacent genes should interact is convenience: adjacent genes can share regulation and expression systems, which can be an evolutionary advantage if their products interact (Teichmann et al, 2001, and refs therein).

In a similar vein, Enright et al and Marcotte et al have both published methods that infer protein-protein interactions from gene fusion events (Enright et al, 1999) (Marcotte et al, 1999). These methods rely on the domain-swap model of evolution (section 1.4.2.2). They predict that if domains A and B are fused in one protein, then whenever orthologues (ie, homologues of the same function) of A and B are on separate proteins, these proteins will interact. Although these methods have enjoyed some success, in their current form they inevitably suffer from high rates of overprediction.

1.5.1.2 Phylogenetic profiles

Phylogenetic profiles are based on the following hypothesis: if two proteins A and B are functionally linked, then a given genome will have either both A and B or neither one. In the method of Pellegrini et al (Pellegrini et al, 1999), two proteins are predicted to interact, directly or indirectly, if their pattern of occurrence among genomes is similar. This method is better at detecting pairs of proteins that share a function or metabolic pathway than those that interact directly. In an assessment by Huynen et al (Huynen et al, 2000), 34% of pairs predicted for *Mycoplasma genitalium* were functionally linked. Because phylogenetic profiles draw their strength from the sequencing of whole genomes and the ability to detect orthologues across them, it is likely to become more powerful with time.

1.5.2 Structure-level methods

All sequence-based methods to localize molecular interactions are relatively new. Prior to the beginning of this work, only a handful of methods for analysis or prediction had been described. Four are summarized below. All rely on the general premise that restricted evolutionary variability of residues reflects their functional importance. Or more specifically, if a protein-protein interaction plays an important functional role, it is interesting to study how patterns of evolutionary conservation in the protomer sequences relate to the maintenance of the interaction. Most also benefit from the fact that residues usually involved in binding are on the molecular surface and surface conservation is generally low. This potentially high signal to noise ratio arises because changes in surface residues do not generally influence folding and overall stability as much as changes in residues at the structural core, so any mutational intolerance that does exist can be detected more easily.

1.5.2.1 Correlated mutations

Consider a pair of residues whose interaction (or non-interaction) is essential for the host's survival. If one residue mutates, that interaction may be lost and the host's fitness severely impaired. However, if the other residue undergoes a compensatory mutation, either at the same time or a few generations later, the interaction, and the host's fitness, could be preserved. Compensatory mutations, usually detected as correlated mutations, are thus considered evidence that two residues are functionally linked. Correlated mutations within a protein sequence have been shown to be important in maintaining stability and function (Serrano et al, 1990) (Shindyalov et al, 1994) (Taylor & Hatrick, 1994) (Göbel et al, 1994).

Pazos et al (Pazos et al, 1997) took this further. They proposed correlated mutations should occur at domain-domain interfaces and, if they did, they should also occur at protein-protein interfaces. Identifying correlations within a single domain sequence is more straightforward than in either a multidomain sequence or between different proteins. In order to identify correlated residue pairs, the residues involved must have a long and intimately linked evolutionary history. Ideally, they should also have been subject to a similar rate of evolution. Residues from different proteins rarely meet these criteria. Even in multidomain sequences, which have often arisen from recent domain-insertion, splicing or recombination events, a linked evolutionary history between residues is not guaranteed. Moreover, it is unwise to assume domain-domain interfaces are under the same kind of evolutionary pressure borne by protein-protein interfaces. After all, two proteins can have the choice not to interact.

Pazos et al performed their analysis on 21 two-domain proteins and one dimer. For each protein, they calculated correlations of amino acid substitutions for all residue pairs in a multiple sequence alignment. Then, to test the hypothesis that correlated mutations identified residues in contact, they compared the pairwise correlations with pairwise residue distances in the crystal structure. Unfortunately, their results only weakly supported their hypothesis and their method as it stood was not obviously useful for predicting protein-protein interactions.

Lockless & Ranganathan (Lockless & Ranganathan, 1999) used correlated mutations, calculated differently from Pazos et al, to test their hypothesis that residues at protein-protein binding sites are energetically coupled with residues distant from the interface. These distal residues support binding indirectly, stabilizing association by allowing binding energy to propagate through the structure rather than being localized at the interface. Their results, performed on just one complex but corroborated by mutagenesis experiments, showed that correlated mutations can correspond well to energetic couplings.

1.5.2.2 Evolutionary tracing at binding site

Lichtarge et al (Lichtarge et al, 1996) described a method for defining functional residues at binding sites. They start with the multiple alignment and associated phylogenetic tree for a protein of interest. Their "evolutionary trace" method then defines a series of cross sections of the tree. A cross section near the root defines a few major branches, each of which represents a low-resolution division of the families. Moving the cross section away from the root and towards the leaves, subfamilies are defined that are progressively smaller and more numerous. This is done in discrete stages to produce a subfamily grouping for each level of sequence identity. At each stage, aligned residues are classified as "conserved", ie, invariant through the entire alignment, "class-specific", ie, invariant within a clade, or "neutral", ie, variable within the clade. As the families become smaller, so does the sequence variation they describe, resulting in the number of class-specific residues increasing with resolution.

Lichtarge et al used an interactive molecular graphics program to view the protein of interest, colouring-in conserved and class-specific residues identified for increasing levels of sequence identity resolution. They found the patterns of clade-specific residue conservation correlated well with observed patterns of relative binding energy. Their method was not, however, particularly automatic. Nor was it predictive. Rather it was a form of exploratory data analysis; after all, considering residue conservation at multiple cross-sections of the family tree is a complex business whose interpretation is ultimately subjective.

1.5.2.3 Searching for 3D binding motifs

De Rinaldis et al (de Rinaldis et al, 1998) developed a search tool that uses evolutionary information to compare binding motifs on protein surfaces. Given a query structure and its sequence alignment, their method maps amino acid substitution patterns of surface residues onto a three-dimensional grid. The grid, which corresponds to a coarse grain model of the structure, is filtered to remove unconserved positions and acts as a profile of the surface. For each protein in a database of structures, they then compare the amino acids on its surface with distributions in the profile and assess the similarity of any aligned residues using amino acid exchange probabilities from a mutation data matrix. Their method could be used to identify homologous surface patches involved in interfaces.

1.6 Conservation at protein-protein interfaces

Intuitively, protein-protein interfaces should be conserved among similarly interacting orthologues. After all, if a protein's function is common within a homologous family and essential or advantageous for the survival of the host organism, the maintenance of that function describes the limits to which mutational variation in the sequence may be tolerated. Moreover, if protein-protein interfaces are conserved, then conservation could be used predictively, discriminating spurious interfaces from true biological ones.

This thesis explores the above themes. It breaks down into three main work chapters. Chapter 2 examines the challenges involved in extracting quantitative evolutionary information from multiple sequence alignments. It surveys the range of strategies that have been used to score residue conservation and develops a new conservation score. Chapter 3 tests the premise that protein-protein interfaces are conserved. Rigorous statistics and the conservation score of chapter 2 are used in this analysis of a small data set of manually validated homodimers. Chapter 4 develops the analysis methods of chapter 3 into a predictive method. This chapter assesses the utility of conservation and size in discriminating biological from nonbiological crystal contacts.

Chapter 5 concludes this work, consolidating the results and suggests new directions for this fertile area.

Chapter 2

Scoring residue conservation

2.1 Introduction

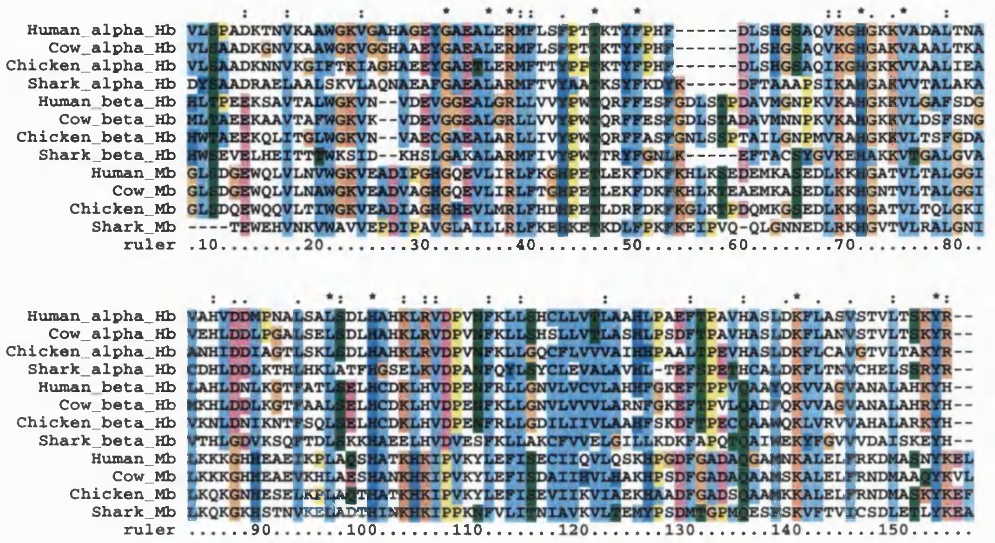
A multiple sequence alignment is a historical record. The patterns of amino acid variability in its columns tell a story of evolutionary pressure, mutation, recombination and genetic drift that often spans many millions of years. This story can be read in different ways, depending on which model of evolution is deemed most appropriate.

According to the neutral model of molecular evolution (Page & Holmes, 1998, and refs therein), once a protein has evolved to a useful level of functionality, most new mutations are either deleterious, in which case they are removed by negative selection, or neutral, in which case they are kept. Most of the substitutions observed in an alignment are therefore neutral; rather than representing improvements in a protein, they indicate how tolerant the protein is to change at that position. In an already optimized protein, the rate of substitution will be inversely correlated with the functional constraints acting on that protein. Fibrinopeptides are under fewer functional constraints than ubiquitin; they also evolve about 900 times faster. The most functionally important residues of haemoglobin (see Figure 2.1), those that secure the haem group, show a much lower rate of substitution than others do in the protein.

The selectionist model of molecular evolution offers a different perspective (Page & Holmes, 1998, and refs therein) (see Figure 2.2). It agrees with the neutralist model that most mutations are deleterious and removed by negative selection, but disagrees about those mutations that are kept. According to this model, the majority of accepted mutations confer a selective advantage whereas neutral mutations are rare. The relative merits of the two models, or their compromise (the "nearly neutral" model), are not considered here. Rather, we consider alignments from the perspective of the neutral model only. This model accords better with the idea of conservation among orthologous sequences and is arguably the more evident in alignments from structural biology.

So if the degree of functional constraint dictates how conserved a position is, then the converse must also be true, ie, the degree of conservation must indicate the functional importance of that position. Identifying conserved regions of a protein can be tremendously useful. Residues involved in an active site or a structural core can sometimes be identified with little prior knowledge of the protein structure.

In the past, patterns of conservation in multiple alignments were identified by inspection alone. However, the rapid increase of available sequences and published analyses has emphasized the need for objective, automated methods, and in the last decade or so this has been the subject of considerable research. Much of that work has focused on extracting global patterns and motifs from multiple alignments, often



```
                      :   :      :       *   * *::  .  *   *                 :: *. .*   :
   Human_alpha_Hb VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQVKGHGKKVADALTNA
     Cow_alpha_Hb VLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQVKGHGAKVAAALTKA
 Chicken_alpha_Hb VLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHF------DLSHGSAQIKGHGKKVVAALIEA
   Shark_alpha_Hb DYSAADRAELAALSKVLAQNAEAFGAEALARMFTVYAATKSYFKDYK-----DFTAAAPSIKAHGAKVVTALAKA
    Human_beta_Hb HLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDG
      Cow_beta_Hb MLTAEEKAAVTAFWGKVK--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKVLDSFSNG
  Chicken_beta_Hb HWTAEEKQLITGLWGKVN--VAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDA
    Shark_beta_Hb HWSEVELHEITTTWKSID--KHSLGAKALARMFIVYPWTTRYFGNLK-----EFTACSYGVKEHAKKVTGALGVA
         Human_Mb GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGI
           Cow_Mb GLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGNTVLTALGGI
       Chicken_Mb GLSDQEWQQVLTIWGKVEADIAGHGHEVLMRLFHDHPETLDRFDKFKGLKTPDQMKGSEDLKKHGATVLTQLGKI
         Shark_Mb ----TEWEHVNKVWAVVEPDIPAVGLAILLRLFKEHKETKDLFPKFKEIPVQ-QLGNNEDLRKHGVTVLRALGNI
            ruler .10........20........30........40........50........60........70........80..

                    : ..    .  *:  *  : ::    :  :       :       :    :   .* .  .   :   *:
   Human_alpha_Hb VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR--
     Cow_alpha_Hb VEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTPAVHASLDKFLANVSTVLTSKYR--
 Chicken_alpha_Hb ANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR--
   Shark_alpha_Hb CDHLDDLKTHLHKLATFHGSELKVDPANFQYLSYCLEVALAVHL-TEFTPEQHCALDKFLTNVCHELSSRYR--
    Human_beta_Hb LAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH--
      Cow_beta_Hb MKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANALAHRYH--
  Chicken_beta_Hb VKNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH--
    Shark_beta_Hb VTHLGDVKSQFTDLSKKHAEELHVDVESFKLLAKCFVVELGILLKDKFAPQTQAIWEKYFGVVVDAISKEYH--
         Human_Mb LKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKEL
           Cow_Mb LKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHPSDFGADAQAAMSKALELFRNDMAAQYKVL
       Chicken_Mb LKQKGNHESELKPLAQTHATKHKIPVKYLEFISEVIIKVIAEKHAADFGADSQAAMKKALELFRNDMASKYKEF
         Shark_Mb LKQKGKHSTNVKELADTHINKHKIPPKNFVLITNIAVKVLTEMYPSDMTGPMQESFSKVFTVICSDLETLYKEA
            ruler ......90.......100.......110.......120.......130.......140.......150......
```

Figure 2.1: Multiple sequence alignment of adult α- and β-haemoglobin (Hb) and myoglobin (Mb) from four vertebrate species.

Amino acids are coloured by their physical and chemical properties according to the scheme of CLUSTALX (Thompson et al, 1997). Stars on the top ruler indicate invariant positions. Note the invariant F at position 50 and H at position 101. These residues both bind the haem group and so are functionally constrained. After ref (Page & Holmes, 1998).

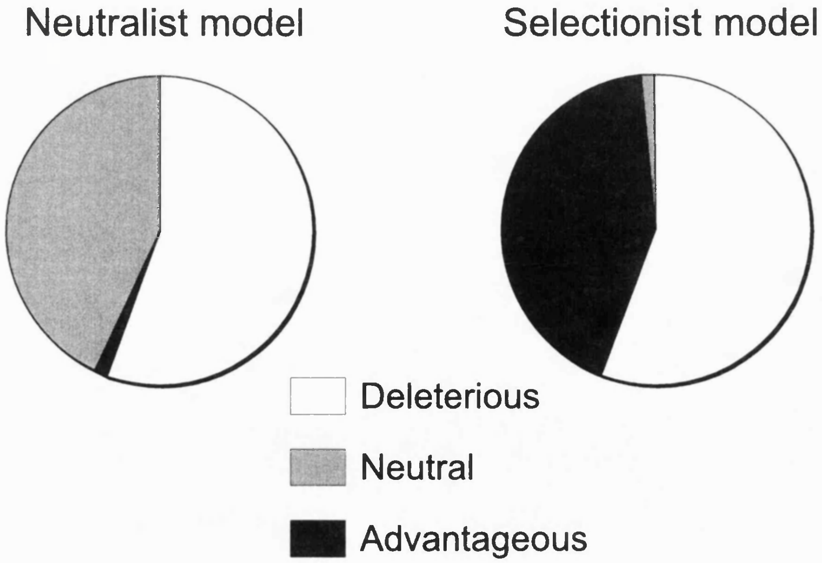


Figure 2.2: Neutralist and selectionist models of molecular evolution.

The pie represents the total number of mutations arising in a gene. The size of each slice represents its contribution to the total according to the neutralist or selectionist model. The slice size is not exact; it merely serves to illustrate the contrasting emphases of the two models. See text for details. After ref (Page & Holmes, 1998).

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
1	D	D	D	D	D	D	I	P	D	L	L
2	D	D	D	D	D	D	I	P	V	L	L
3	D	D	D	D	D	D	I	P	Y	L	L
4	D	D	D	D	D	D	I	P	A	L	L
5	D	D	D	D	D	D	L	W	T		-
6	D	D	E	D	E	E	L	W	K		-
7	D	D	E	D	E	E	L	W	P		-
8	D	D	E	D	E	E	L	W	C		-
9	D	D	E	D	E	F	V	S	R		-
10	D	E	E	F	F	F	V	S	H		-

Figure 2.3: Some example columns from different multiple alignments.
Each labelled column represents a residue position in a multiple sequence alignment. The rows denote the sequence number of a particular amino acid. Amino acids are identified by their one-letter code, gaps by a dash ("-"). Note that column (k) comes from an alignment of ten sequences whereas column (j) comes from an alignment of only four.

with a view to exploring the relationships between homologues and developing diagnostic tests for functions of newly discovered sequences. For instance, statistically robust profile methods, such as PSI-BLAST (Altschul et al, 1997) and those based on hidden Markov models (Eddy, 1996), have become increasingly popular.

Despite these advances there have been few recent insights into the derivation of a quantitative conservation measure for a single aligned position, and there certainly is no standard method. Ask a life scientist how similar two sequences are and he will probably quote a percentage identity or an E-value. Ask him how conserved a position is in a family and the reply is most likely to be qualitative. This chapter discusses what a quantitative measure of conservation should actually measure, and, by surveying almost than twenty scores, examines some of the problems inherent in developing such a score.

2.1.1 Exercises for a conservation score

There is no rigorous mathematical test for judging a conservation measure; if there were, one would use the test and not bother with an additional score. Rather than accuracy then, a conservation score may be judged on its verisimilitude: its ability to depict realism and its concordance with biochemical intuition. Figure 2.3 is part of an attempt to make these abstract notions more concrete. It shows columns of amino acids taken from hypothetical multiple sequence alignments of orthologues. Applying basic biochemical knowledge to this collection of columns reveals some concrete qualitative comparisons. Specifically, from most conserved to least conserved, the following orders seems reasonable: (a) > (b) > (c) > (d) > (e) > (f) , then (g) > (h) > (i), and lastly (j) > (k).

Column (a) contains only D and is therefore the most obviously conserved. Column (b) also contains E, so (b) is more variable than (a). Column (c) contains D and E but is less dominated by any one than (b), so (c) more variable than (b). Column (d) contains nine D and one F; it is clearly more variable than column (a), but is it more variable than column (b)? Phenylalanine is large and non-polar whereas aspartate and glutamate are both smaller and polar. Because the amount of stereochemical variability in column (d) is greater than in column (b) it seems likely a mutation from D to E glutamate would be more tolerable than one from D to F (a conclusion supported by the exchange probabilities in a mutation data matrix; see later). Column (e) implies both conservative substitutions (between D and E) and non-conservative ones (between the acids and phenylalanine). Column (e) is thus the least conserved so far. Column (f) contains the same amino acid types as (e). However, because it is less skewed towards an abundance of D and E, ie, more evenly mixed, (f) is more variable.

Columns (g) and (h) are equivalent in terms of the number and frequency of their amino acids. However, because (g) contains only branch-chain amino acids whereas (h) encompasses a broader mix of stereochemical characteristics, (g) is more suggestive of conservative substitutions in response to negative selective pressure. Column (i) is the most variable column encountered so far, as judged by biochemistry or amino acid frequency.

Columns (j) and (k) illustrate the importance of gaps. Column (j) is taken from an alignment of four sequences. In each sequence, a leucine is present at that position. Column (k) also contains four leucines but, because it comes from an alignment of ten sequences, it also contains six gaps. For column (k), then, there is strong evidence that leucine is not functionally constrained. After all, this amino acid has been shed from six other orthologues with apparent impunity. There is no such evidence for column (j), the conservation of which remains untarnished. The comparison between columns (j) and (k) also highlights the dangers of having too small an alignment. The alignment of (j) could be the same as that of (k) but with six sequences missing; an example of lack of data producing completely different conclusions about the same site.

Figure 2.3 will be used here as a testing ground for some of the scores surveyed in the forthcoming section 2.2.

2.1.2 Requirements of a conservation score

A score that quantifies the degree of conservation at an aligned position should fulfil the following criteria.

1. Mathematical properties. The score should be a function that maps a set of arguments (the input space), which includes the aligned column and possibly other information, to a number (the output space). Convenient scores will have an output space that is continuous and bounded.

2. Amino acid frequency. The score should take account of the relative frequencies of amino acids in a column. For instance, using the columns from Figure 2.3, it should reproduce the ranking (a) > (b) > (c) > (e) > (f).

3. Stereochemical properties. The score should recognize conservative replacements and that some substitutions incur more chemical and physical change than others. For instance, it should score column (g) as more conserved than column (h).

4. Gaps. A preponderance of gaps suggests a position can be deleted without significant loss of protein function. The score should therefore penalize such positions and should rank column (j) as more conserved than column (k). An ideal score might also recognize that, in terms of protein structure, the difference between a small residue, eg, glycine, and a gap is less than between a large residue, eg, tryptophan, and a gap.

5. Sequence weighting. Sometimes a position appears conserved among a number of sequences not because of functional constraint but because those sequences have not had sufficient evolutionary time to diverge. A typical alignment often includes some sequences that are very closely related to each other. These clusters of highly similar sequences may reflect bias in the sequence databases or result from nature's irregular sampling of the space of acceptable mutations. Either way, such clusters can monopolize alignments, masking important information about allowed variability from

more sparsely represented sequences. For instance, the high frequency of aspartate in column (b) may owe more to the tight homology of sequences 1 to 9 than aspartate being stereochemically preferable in that position. A good conservation score should therefore find some way to normalize against redundancy and bias in the alignment. A facile solution would be simply to remove sequences at a certain level of similarity. However, this is wasteful of what little information these removed sequences could contribute. A good conservation score should therefore find some way to normalize against redundancy and bias in the alignment without loss of evolutionary information.

6. Simplicity (de Bono, 1999). Most scoring methods, from E-values that describe sequence similarity to A-level grades, have their limitations. Understanding the shortcomings of these methods is key to employing them wisely and interpreting their results meaningfully. Therefore, on the reasonable assumption that no method is perfect, a good conservation score should be no more complex than it needs to be so its deficits can be understood. To quote Einstein, "everything should be made as simple as possible, but no simpler".

2.2 A survey of conservation scores

Over the last thirty years a number of methods have been proposed to score residue conservation. The scores surveyed here, which number more than fifteen, are presented in approximately increasing order of sophistication in terms of what they try to achieve. For clarity, the names given to each score by its authors are ignored in favour of the following convention. Scores whose values increase with increasing conservation are denoted C_{name}, where the subscript identifies the author. Scores that do the converse are denoted V_{name}.

2.2.1 Symbol frequency scores

Scores in this category consider amino acids as symbols in a uniformly diverse alphabet. They focus on their relative frequency of these symbols and do not account for sequence redundancy in the alignment. Because, by definition, none model stereochemical properties (criterion 3) or weight their sequences (criterion 5) the discussion instead concentrates on how well they fulfil the remaining criteria.

In 1970, Wu & Kabat (Wu & Kabat, 1970) introduced the first widely accepted measure of conservation. Their score, which they used to identify the variable regions on antibodies, was defined as

$$V_{Kabat} = \frac{k}{n_1} \quad \times N ,$$

where k is the number of amino types present at the aligned position, n_1 is the number of times the most commonly occurring amino acid appears there and N is the number of sequences in the alignment. The variable N acts as a scaling factor and is constant for a given alignment. For clarity, this survey will tend to set such constants apart from the main equation.

Applying the score to Figure 2.3, V_{Kabat} correctly reproduces the ranks (a) > (b) > (c) > (e) but fails to distinguish (e) from (f). This is because it cares only about the frequency of the most commonly occurring symbol and ignores the frequencies of the rest. V_{Kabat} has other problems; for one, it is discontinuous along its output space. A strictly conserved column, such as column (a), always scores 1. A column that is strictly conserved except for one aberrant amino acid is $2\frac{N}{N-1} > 2$, regardless of how many sequences are in the alignment. This discontinuity is biologically meaningless (Shenkin et al, 1991). The score also fails to consider gaps, and so fulfils only the criterion of simplicity.

Jores et al (Jores et al, 1990) recognized the Kabat score's inability to distinguish (e) from (f) and in response proposed a modified version:

$$V_{Jores} = \frac{k_{pair}}{n_{pair_1}} \times \frac{1}{2}N(N-1)\,,$$

where $\frac{1}{2}N(N-1)$ is the number of possible pairs of amino acids in the column, k_{pair} is the number of distinct pairs and n_{pair_1} is the number times the most frequently distinct pair occurs. By considering pairs rather than singlets, this score improves upon V_{Kabat}. However, all the other deficits remain. It is still discontinuous: whereas complete conservation scores one, the next most conserved value possible is $2\,\frac{N}{N-2} > 2$. It does not account for gaps. Even its simplicity is questionable: V_{Jores} does the same job, but is significantly more awkward to compute, than the symbol entropy scores discussed later.

Lockless & Ranganathan (Lockless & Ranganathan, 1999) propose a different type of symbol frequency score. They measure the conservation at an aligned position as the extent to which amino acid frequencies at that position deviate from frequencies over the whole alignment. To model this deviation, they employ binomial probabilities. If an amino acid a occurs in the sequence databases at fractional frequency q_a, then the probability of a occurring n_a times in a column of N residues is $P(X = n_a)$ where $X \sim Bin(N, q_a)$. For example, if half the amino acids in SWISSPROT (Bairoch & Apweiler, 2000) were Ds, then the probability of D occurring nine times in column of ten residues is the same as the probability of getting nine heads from ten coin tosses. This probability compared with the probability for the overall frequency of a in the alignment to give a measure of deviation

$$d\,(n_a, \bar{n}_a) = \ln\left(\frac{P\,(X = n_a)}{P\,(X = \bar{n}_a)}\right)\,,$$

where $\bar{n}_a$ is the average frequency of a in the whole alignment. The distance d describes how much the frequency a at the position differs from that of a across the alignment. When these frequencies are the same, $d = 0$, when they are different d may be positive or negative. The conservation for the column, $C_{Lockless}$[1], is taken as the root mean square deviation over all 20 amino acids, ie,

$$C_{Lockless} = \sqrt{\sum_a d\,(n_a, \bar{n}_a)^2}\,.$$

If a single column can be represented by a point in 20-dimensional space of binomial probabilities, then $C_{Lockless}$ measures the Euclidean distance between that point and the point representing the "average" column.

In a typically diverse alignment, Lockless & Ranganathan's score identifies columns dominated by only a few amino acids, since the binomial probabilities of these columns would be small. Some strictly conserved columns score higher than others. For instance, if cysteine occurs least frequently in the alignment, a strictly conserved column of cysteine will score higher than a strictly conserved column of histidine. Although this has some intuitive appeal – strictly conserved columns of rare amino acids are visually more striking in an alignment – the authors do not argue its case. But this arbitrariness is symptomatic of a deeper malaise: that $C_{Lockless}$ is complex. Its purpose is to measure how different a column is from the rest of the

[1] In their original paper, Lockless & Ranganathan presented their score as $\Delta G_i^{stat} = kT^* \sqrt{\sum_a \left(\ln \frac{P(X=n_a)}{P(X=\bar{n}_a)}\right)^2}$. For clarity, gratuitous references to thermodynamics are removed to give $C_{Lockless}$.

alignment. However, d could be calculated far more simply, say, as the Euclidean distance between the two sets of amino acid frequencies. Instead $C_{Lockless}$ uses a binomial model that brings in further data, namely the frequencies of amino acids from a sequence database. This extra information adds considerably complexity to the score, but is this complexity worth it? Considering that (arguably) more important information about stereochemistry, gaps, and the like is omitted, it seems not.

2.2.2 Symbol entropy scores

Symbol entropy scores are a specialization of symbol frequency scores (section 2.2.1). Scores in this category all account for the relative frequencies of symbols using Shannon's entropy or variations thereof.

2.2.2.1 Background

Shannon's information theoretic entropy (Shannon, 1948) (hereafter referred to as "Shannon's entropy") is an often-used measure of diversity (Baczkowski et al, 1997) (Durbin et al, 1998). It can be derived from two roots, one combinatoric and one information theoretic. The combinatoric derivation goes as follows. Given 10 coloured balls, of which 5 are red, 2 green and 3 yellow, the number of distinct sequences you can make is $10!/(5!2!3!) = 2520$. More generally, given N objects that fall into K types, the number of distinct ways they can be permuted is given by the multinomial coefficient,

$$W = \frac{N!}{\prod_{i=1}^{K} n_i}, \tag{2.1}$$

where n_i is the frequency of the ith type. As N becomes large, $N!$ can be calculated using Sterling's approximation, $\ln N! \sim N \ln N - N$, such that

$$\ln W = -N \sum_{i}^{K} p_i \ln p_i ,$$

where $p_i = n_i/N$, the fractional frequency of type i. Transforming linearly gives the Shannon entropy:

$$S = -\sum_{i}^{K} p_i \log_2 p_i . \tag{2.2}$$

The quantities S and W monotonically increase with each other. S ranges from zero, when objects of only one type are present, to $S_{max} = \log_2 K \geq 0$, when all types are present in equal proportion. It has been shown that Shannon's entropy belongs to a general class of diversity index (Good, 1953),

$$D(\alpha, \beta) = \sum_{i}^{K} p_i^{\alpha} (-\log p_i)^{\beta} .$$

Of this class, both Shannon's entropy ($D(1,1)$) and Simpson's index ($D(2,0) = \sum_i^K p_i^2$) have been used in ecology for measuring species diversity (Baczkowski et al, 1997, and refs therein). Note that the base of the logarithm affects only the unit of measurement, not the score itself since, for any a and b, $\log_a x \propto \log_b x$.

The original use of Shannon's entropy was in information theory, a branch of electronic engineering that examines communication and the handling of information (Gregory & Zangwill, 1987, and refs therein) (Durbin et al, 1998). In many older telecommunication systems, such as radio, the signal constructs the output. In more modern systems, such as teleprinting and many digital systems, the range of possible outputs is small and known in advance. This allows a much more economical approach called encoding, in which the signal selects output from a finite list. The selective information content of an encoded signal depends not on the size or complexity of the output as such, but on the number of alternative forms it might have taken, and on the relative likelihood of each. The simplest selective operation is one that chooses between two equally likely possibilities, eg, the symbols A and B. This is a binary decision and the gain in information when it is made is one binary digit (bit). Choosing between four symbols requires two bits, eg, to identify one of {A1, A2, B1, B2}, you must make two binary decisions: "A or B?" and "1 or 2?". More generally, the number of binary decisions needed to choose between K equiprobable symbols is $\log_2 K$. Rearranging gives

$$S = \log_2 K = -\log_2 \frac{1}{K} = -\log_2 p\,,$$

where p is the probability of selecting any one symbol. In this context, S is the information required to make the selection and hence is a measure of uncertainty. If symbol A is far more likely to be selected than B, then the outcome of the selection is more certain. This can be accommodated by partitioning the $-\log p$ with prior probabilities:

$$S = p_A\,(-\log_2 p_A) + p_B\,(-\log_2 p_B)\,,$$

where p_A and p_B are the probabilities of selecting A and B respectively. Generalizing for K symbols gives the Shannon entropy (equation 2.2). The total selective information content of a signal is defined as the amount of uncertainty it resolves:

$$I = S_{\text{before}} - S_{\text{after}}\,, \tag{2.3}$$

ie, the difference between the information entropy before the signal and after it.

2.2.2.2 Scores

In 1991, two groups proposed residue conservation scores based on Shannon's entropy. Until then, entropy had been used for scoring positional conservation, but only in nucleotide sequences (Schneider, 1997, and refs therein). Sander & Schneider (Sander & Schneider, 1991) defined their score as a normalized Shannon's entropy:

$$V_{Schneider} = -\sum_i^K p_i \ln p_i \quad \times \frac{1}{\ln K}\,,$$

where $K = 20$, representing the 20 amino acid types. Shenkin et al (Shenkin et al, 1991) proposed the related score

$$V_{Shenkin} = 2^S \quad \times 6,$$

where S is Shannon's entropy 2.2 and K in that equation is also 20. These scores are transformations of each other and so are trivially different. Both purport to have conveniently bounded ranges: $0 \leq V_{Schneider} \leq 1$ and $6 \leq V_{Shenkin} \leq 120$. However, neither would score column (i) in Figure 2.3 as maximally variable. This is more a minor artifact than a serious deficit. Shannon scores treat columns of residues as if they were rows of coloured balls. Maximal diversity occurs when all colours are represented evenly. But if there are more colours than there are balls to represent them, this limit can never be reached. Similarly, $V_{Schneider}$ and $V_{Shenkin}$ can reach their top value only when there are at least 20 sequences in the alignment.

Gerstein & Altman (Gerstein & Altman, 1995) present another variation on this theme. To compare sequence conservation with structural conservation in a multiple alignment of protein structures, they define

$V_{Gerstein}$, which measures the entropy of a position relative to that if the sequences were aligned randomly:

$$V_{Gerstein} = \sum_{i}^{K} \overline{p}_i \log_2 \overline{p}_i - \sum_{i}^{K} p_i \log_2 p_i \,,$$

where $\overline{p}_i$ is the average frequency of amino acid i in the alignment and $K = 20$. This score, which is in the same form as equation 2.3, measures the information content of the position in bits. Not that this bestows any particular advantage; like the other entropy scores, $V_{Gerstein}$ delivers nothing grander than a conveniently expressed multinomial coefficient (equation 2.1).

Like V_{Jores}, Shannon-based scores rank (a), (b), (c), (e) and (f) correctly. Unlike V_{Jores}, they are continuous. In a column strictly conserved but for one aberrant residue, the entropy decreases to the score's minimum with an increasing number of sequences. Shannon's entropy is also much simpler to calculate: whereas V_{Jores} requires information about pair frequencies, which itself requires combinatoric calculations, entropy requires only fractional frequencies of the symbol types; the entropy equation does the combinatorics.

So the symbol entropy scores fulfil criteria of mathematical properties and amino acid frequency, and, with their straightforward calculation, acquit themselves of complexity. But as well as being simple, these scores are simplistic. Amino acids are not coloured balls, no matter how mathematically convenient it is to think otherwise. None of these scores could distinguish column (g) from (h) in Figure 2.3. When Gerstein & Altman compare structural conservation, using an atom coordinates-based scheme, with sequence conservation, using $V_{Gerstein}$, they find the two have little in common (Gerstein & Altman, 1995). Perhaps a sequence conservation score that considered stereochemistry would have led them to a different conclusion.[2]

More worryingly, none of these scores account for gaps. This is a problem. In the Shannon scheme, it is most natural to consider a gap as another symbol type, the "21st" amino acid. Doing this, however, has absurd consequences. For instance, column (k), which is predominantly gapped, would score as more conserved than columns (c) or (g).

2.2.3 Stereochemical property scores

Scores in this category consider only the stereochemical properties of the amino acids in a column. These scores typify a view orthogonal to that of the symbol frequency and symbol entropy scores described above.

In 1986, Taylor (Taylor, 1986) classified amino acid types according to their stereochemical properties and their patterns of conservation in the Dayhoff mutation data matrix (Dayhoff et al, 1978). He embodied this consolidation of mutational and physical data in a Venn diagram (Figure 2.4), in which each overlapping set represents a distinct physical or chemical property. Taylor then devised a set theoretic method based on this diagram to score positional conservation. His method finds the smallest set or subset that describes the amino acid types observed at an aligned position. The variability of the column is taken as the total number of residue types belonging to that set. The number of possible subsets of the Venn diagram is large and many of these sets have little physical meaning. To reduce the possibility of high conservation being ascribed to meaningless subsets, Taylor compiled a list of 70 sets and subsets that might reasonably be

[2] Interestingly, an almost identical criticism has recently been leveled by Mirny & Shakhnovich (Mirny & Shakhnovich, 2001) at a comparison of structure and sequence conservation by Plaxco et al (Plaxco et al, 2000). Plaxco et al used a score much like $V_{Gerstein}$ and formed similarly heretical conclusions.

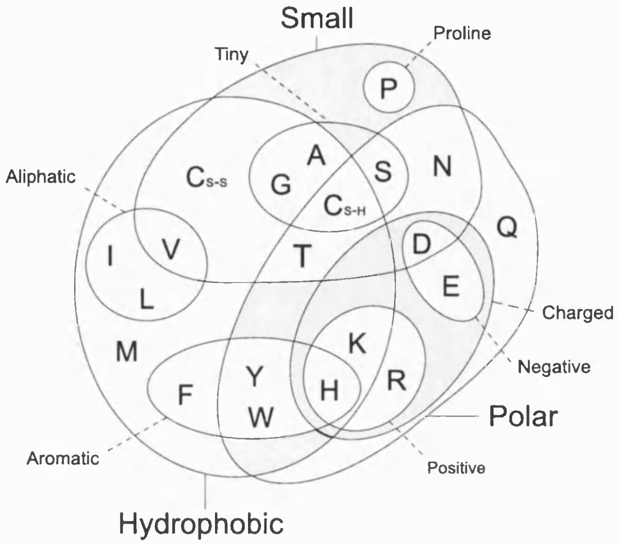


Figure 2.4: Taylor's Venn diagram of amino acid properties.
Taylor argues Cys should appear twice because although the reduced form (C_{S-H}) has similar properties to Ser, the oxidized form (C_{S-S}) is more like Val. Adapted from refs (Taylor, 1986) and (Livingstone & Barton, 1993).

conserved, and suggested only these "valid" sets should be considered. Taylor's score can be expressed as

$$V_{Taylor} = \min\left(n\left(\{X : Aligned \subseteq X \text{ and} X \in Valid\}\right)\right) ,$$

where *Aligned* is the set of amino acids at the aligned position, *Valid* is Taylor's set of 70 valid sets and $n(X)$ is the number of elements in set X. V_{Taylor} ranges from 1 to 20.

Taylor's score accomplishes some things the symbol scores could not. It recognizes that column (b) from Figure 2.3 is more conserved than (d) and that (g) is more conserved than (i). It does not explicitly model gaps but there is a natural way these could be incorporated into the scheme: a gap could belong only to the largest superset. But Taylor's score is clumsy. The ad hoc clause of reducing the number of valid sets to 70 makes the score more computationally tractable but diminishes its simplicity and elegance. To interpret the score properly one must accept that some subsets in the Venn diagram are forbidden. This introduces a degree of subjectivity on top of that supplied by the Venn diagram itself.

Taylor's score has more conspicuous problems. First, the score of strictly conserved columns depends on the amino acid: a column of Ps scores 1, Hs score 3, Ws score 4. Similarly, column (g) in Figure 2.3 would score the same as a strictly conserved column of I. Second, it fails to account for amino acid frequencies and cannot distinguish column (b) from (c) or (e) from (f).

Clearly, a Venn diagram is picturesque but unwieldy. Could it be abridged to something more convenient? Zvelibil et al (Zvelibil et al, 1987) reduce Taylor's diagram to a truth table of amino acids vs ten property descriptors (Figure 2.5). They define their score as

$$V_{Zvelibil} = n_{const} \quad \times \frac{1}{10} ,$$

where n_{const} is the number of properties whose state (ie, truth or falsehood) is constant for all amino acids in the column. For example, column (b) in Figure 2.3 contains D and E, which share 9 properties, and

	I	L	V	C	A	G	M	F	Y	W	H	K	R	E	Q	D	N	S	T	P	B	Z	X	Δ	
1	●	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○	○	●	○	○	○	●	●	Hydrophobic
2	○	○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	Polar
3	○	○	●	●	●	●	○	○	○	○	○	○	○	○	○	●	●	●	●	●	○	○	●	●	Small
4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○	●	●	Proline
5	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	●	●	Tiny
6	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	Aliphatic
7	○	○	○	○	○	○	○	●	●	●	●	○	○	○	○	○	○	○	○	○	○	○	●	●	Aromatic
8	○	○	○	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	○	○	○	○	●	●	Positive
9	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	●	○	○	○	○	○	○	●	●	Negative
10	○	○	○	○	○	○	○	○	○	○	●	●	●	●	○	●	○	○	○	○	○	○	●	●	Charged

Figure 2.5: Truth table profile of amino acid properties.
Amino acid (across) are each described in terms of ten properties (down). A filled circle means the amino acid above it possesses that property. The symbol "Δ" represents a gap, which is considered to have all properties. Adapted from ref (Livingstone & Barton, 1993).

scores 0.9. Although it has a less erratic output space, $V_{Zvelibil}$ retains V_{Taylor}'s failure to account for amino acid frequency. In their program AMAS (Analysis of Multiply Aligned Sequences), Livingstone & Barton (Livingstone & Barton, 1993) turn this weakness into a strength. AMAS uses $V_{Zvelibil}$ to split sequences into subgroups and thus infer an evolutionary or functional hierarchy from the alignment. For example, given column (f) of Figure 2.3 AMAS may decide sequences $\{9,10\}$ are in a different subfamily from sequences $\{1..8\}$ because the $V_{Zvelibil}$ score within these sets is much higher than in their superset $\{1..10\}$. The success of AMAS demonstrates $V_{Zvelibil}$ is better suited to this kind of selectionist analysis. In particular, $V_{Zvelibil}$ could add welcome sophistication to the evolutionary trace method of Lichtarge et al (Lichtarge et al, 1996) (see chapter 1.5.2.2).

2.2.4 Mutation data scores

Scores in this category use mutation data from a substitution matrix to quantify stereochemical variability in an aligned column. No scores in this category normalize against sequence redundancy in the alignment.

2.2.4.1 Background

Substitution matrices provide a quantitative and reasonably objective measure of amino acid similarity. A substitution matrix is a table of amino acid exchange probabilities derived from an analysis of the evolutionary changes seen in a group of homologous proteins. Figure 2.6 shows BLOSUM62 (Henikoff & Henikoff, 1992), a popular substitution matrix. Others well known matrices include the Dayhoff Mutation Data Matrix (MDM) (Dayhoff et al, 1978) and the Pairwise Exchange Table (PET) (Jones et al, 1992). The non-diagonal pairwise scores indicate how likely one amino acid is to be substituted by another in a homologous protein. The diagonal scores, which pitch an amino acid against itself, indicate how likely an amino acid is to substituted at all, ie, its "mutability". Because the chance of a substitution increases with evolutionary time, any particular matrix is parameterized by some kind of evolutionary distance. For instance, BLOSUM62 captures the rates of exchange one would expect in homologues that were 62% identical. Evolutionary distance not only affects the likelihood of a mutation but also its nature. Mutations that differentiate close homologues are mostly influenced by the genetic code, whereas those separating divergent sequences are dominated by stereochemistry (Benner et al, 1994).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	**7**	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**

Figure 2.6: The BLOSUM62 substitution matrix.
BLOSUM62 (Henikoff & Henikoff, 1992), which stands for BLOcks SUbstitution Matrices from sequences clustered at 62% identity, is constructed from BLOCKS (Henikoff & Henikoff, 1991) version 5.0 and SWISSPROT (Bairoch & Apweiler, 2000) version 22. Entries for identical residue are on the diagonal and highlighted in bold.

The primary purpose of substitution matrices is to help evaluate whether an observed alignment of two sequences, or two residues, is biologically correct or an artifact of the alignment algorithm. This purpose is evident in their construction. The non-diagonal elements of the matrix describe the likelihood of a protein substituting amino acid a for b as a ratio of two probabilities

$$R(a,b) = \frac{P(a,b\,|Match)}{P(a,b\,|Random)}, \tag{2.4}$$

where $P(a,b\,|Match)$ is the probability of a substituting for b under the assumption their positions are biologically equivalent in the protein, and $P(a,b\,|Random)$ is the probability of observing a and b aligned randomly, which is a function of their respective overall frequencies in the database. Popular pairwise alignment algorithms score a given alignment by accumulating $R(a,b)$ over every position in the sequence, typically in conjunction with a length-dependent gap penalty. This accumulation is awkward with raw $R(a,b)$ because it involves many slow floating-point multiplications. For computational convenience then, $R(a,b)$ is instead expressed as its logarithm, scaled and rounded to the nearest integer:

$$m(a,b) = \mathrm{int}[\lambda \log R(a,b)],$$

where λ is a scaling constant. Probabilities may now be accumulated by simply summing $m(a,b)$ (Durbin et al, 1998). The likelihood ratio 2.4 may or may not be used to calculate diagonal elements, depending on the matrix. If not, as is the case for the PET, then $m(a,a)$ is derived from the observed mutability of a in the dataset used to compile the matrix. Simplistically, the observed mutability of a in a trusted alignment of two sequences is the number of times a is seen to change divided by the number of times a occurs. This mutability, calculated over the whole dataset, is typically normalized in several ways before it reaches the substitution matrix. Most significantly, it is inverted such that a large $m(a,a)$ indicates a is unlikely to mutate within the prescribed evolutionary time.

But if the diagonal scores in the matrix measure the inertia of an amino acid against mutation, then this is where the analogy between a substitution matrix and a similarity matrix breaks down. The diagonal of a substitution matrix helps an alignment algorithm decide whether two amino acids should be aligned. Its use is normative concerning the alignment. A similarity matrix used by a conservation score must assess the similarity of the amino acids in a column. Its use is descriptive concerning the alignment. The conservation score does not seek to question the validity of the alignment; rather, it assumes the alignment is correct and seeks to describe its features. These two motives are fundamentally different. For instance, the diagonal in a substitution matrix tells you that Trp is rarely substituted whereas Arg is substituted more readily. This makes sense; Trp is unique among amino acids whereas Arg has more obvious replacements. A conservation score that used the substitution matrix to measure similarity would therefore rate a column containing only Trp as more conserved than one containing only Arg. This is would be wrong. Given that we trust the alignment, strict conservation of a more replaceable amino acid suggests a greater evolutionary constraint on that position. The functional constraint could be such that, although other amino acids are similar, because they differ even slightly in their geometry and chemistry, being similar is not enough. Therefore, a measure of replaceability is not just different from a measure of similarity, it is actually at odds with a descriptive measure of conservation.

How can this be resolved? The simplest answer is to redefine the diagonal, hence explicitly converting the substitution matrix into a similarity matrix, ideally in a way that minimally disturbs the off-diagonal values. For example, all diagonal values could be constant, set to the highest diagonal or off-diagonal value. Alternatively, the entire matrix could be normalized to take into account diagonal values – though this would count as perturbation of perfectly good off-diagonal values. If the similarity matrix is explicitly for measuring conservation, there is even a case for scaling diagonal values inversely to their values in the substitution matrix; ie, the more replaceable an amino acid is, the more significant an event is its conservation and so the higher its self-similarity score.

2.2.4.2 Scores

Karlin & Brocchieri (Karlin & Brocchieri, 1996) propose the following score, which they use to study conserved positions in DNA-binding proteins:

$$C_{Karlin}(x) = \sum_{i}^{N} \sum_{j>i}^{N} M(s_i(x), s_j(x)) \quad \times \frac{2}{N(N-1)},$$

where $s_i(x)$ is the amino acid at column x in the ith sequence, and $M(a,b)$ is the similarity between amino acids a and b. The similarity matrix M is defined such that

$$M(a,b) = \frac{m(a,b)}{\sqrt{m(a,a)\, m(b,b)}}, \tag{2.5}$$

where m is BLOSUM62 or a similar substitution matrix. The normalization 2.5 ensures that $M(a,a) = 1$ always and that, provided m has a typical range, $-1 < M(a,b) \leq 1$. This in turn means C_{Karlin} ranges from -1 to 1. C_{Karlin} is a so-called "sum of pairs" (SP) score. It describes conservation by calculating the sum of all possible pairwise similarities between residues in an aligned column. For example, column (e) in Figure 2.3 contains five Ds, four Es and one F. This would be described by the number

$$C_{Karlin} = \frac{10M(\mathtt{D},\mathtt{D}) + 6M(\mathtt{E},\mathtt{E}) + 20M(\mathtt{D},\mathtt{E}) + 5M(\mathtt{D},\mathtt{F}) + 4M(\mathtt{E},\mathtt{F})}{10+6+20+5+4}.$$

From the above calculation, it is clear that this score reflects not only stereochemical variation in a column but also the relative frequencies of the amino acids present. If the distribution of D, E and F was even more dominated by D, there would be more high scoring $M(\mathtt{D},\mathtt{D})$ terms and fewer low scoring terms.

One criticism leveled at SP scores is that they do not make sense in terms of what the statistic $m(a,b)$ means (Durbin et al, 1998). Nature is parsimonious and it is improbable that the diversity observed in column (e) of Figure 2.3 truly results from $20+5+4=29$ amino acid substitutions among 10 homologues. However, if one treats the substitution matrix as no more than a quantitative guide to pairwise amino acid similarity, the SP score is no less than a convenient way to consolidate this two-dimensional information into a single number. Besides, the perceived over-counting of amino acid substitutions in the SP scores is assuaged somewhat by its over-counting of self-similarity terms. In fact, SP scores can be seen as a tug-of-war between self-similarity and substitution, ie:

$$C_{Karlin} \propto \sum_{a}^{K} \sum_{b \geq a}^{K} \begin{cases} a=b & \rightarrow & \frac{n_a(n_a-1)}{2} M(a,a) \\ a \neq b & \rightarrow & n_a n_b M(a,b) \end{cases}, \tag{2.6}$$

where K is the number of amino acid types and n_a is the number of occurrence of amino acid type a. The upper term ($a=b$) bestows high-scoring $M(a,a)$ values whereas the lower term provides predominantly low-scoring $M(a,b)$ values. But even if it escapes this criticism, C_{Karlin} deserves a further reproach: it does not account for gaps.

The score of Armon et al (Armon et al, 2001) does account for gaps. Armon et al present "ConSurf", an implementation and extension of the evolutionary trace method of Lichtarge et al (Lichtarge et al, 1996). ConSurf measures conservation using a variation on the SP theme, defining its score as

$$V_{Armon} = \sum_{a>b}^{20} f_{ab} D(a,b),$$

where

$$f_{ab} = \begin{cases} 1 & \text{if amino acids } a \text{ and } b \text{ present} \\ 0 & \text{otherwise} \end{cases},$$

and D is a dissimilarity matrix[3]. In the tug-of-war notation of equation 2.6, this can be expressed as

$$V_{Armon} = \sum_{a}^{K} \sum_{b \geq a}^{K} \begin{cases} a=b & \rightarrow & \begin{cases} n_a > 1 & \rightarrow & n_a D(a,a) \\ n_a \leq 1 & \rightarrow & 0 \end{cases} \\ a \neq b & \rightarrow & (n_a + n_b) D(a,b) \end{cases}, \tag{2.7}$$

where the upper terms contribute conserved scores and the lower terms contribute variability. Rather than basing their similarities on a substitution matrix, Armon et al use a physico-chemical distance matrix

[3] Technically, f_{ab} is the number of times a and b are seen to exchange in a phylogenetic tree of the sequences. However, the definition above is a fair approximation (Nir Ben-Tal, personal correspondence).

(Miyata et al, 1979)[4]. The physico-chemical distance is defined as

$$D(a,b) = \sqrt{\left(\frac{\Delta\text{pol}_{ab}}{\sigma_{\text{pol}}}\right)^2 + \left(\frac{\Delta\text{vol}_{ab}}{\sigma_{\text{vol}}}\right)^2},$$

where Δpol_{ab} and Δvol_{ab} are the differences in polarity and volume between amino acids a and b, and σ_{pol} and σ_{vol} are the standard deviations for these properties. D is zero for all $D(a,a)$, as much as 4.88 (=$D(\mathtt{D},\mathtt{W})$) for comparisons between real amino acids, 6 for $D(a,-)$ and 0.5 for $D(-,-)$. Distances for gaps are set heuristically.

Gaps aside, V_{Armon} and C_{Karlin} are only subtly different. C_{Karlin} emphasizes self-similarity more than V_{Armon}. This is evident from their self-similar terms in equations 2.6 and 2.7: in C_{Karlin}, the coefficient of $M(a,a)$ grows quadratically with respect n_a; for V_{Armon}, the coefficient of $D(a,a)$ grows linearly. Their substitutions terms also differ, but do not correct the imbalance. For instance, applying V_{Armon} to column (e) of Figure 2.3 produces the number

$$V_{Armon} = \frac{5D(\mathtt{D},\mathtt{D}) + 4(\mathtt{E},\mathtt{E}) + 9D(\mathtt{D},\mathtt{E}) + 6D(\mathtt{D},\mathtt{F}) + 5D(\mathtt{E},\mathtt{F})}{5+4+9+6+5}. \tag{2.8}$$

Comparing this with calculation 2.8 underlines this difference in emphasis: in C_{Karlin} $16/29 = 0.55$ of the terms are self-similar whereas in V_{Armon} this fraction is lower at $9/20 = 0.45$. The scores are also different in that V_{Armon} employs a physico-chemical distance matrix. But to what advantage? Armon et al argue that polarity and volume are the most important factors governing conservation of amino acid type. The most obvious way to check this would be to see if these factors dominate a substitution matrix. If volume and polarity do dominate, Armon et al might as well have used a substitution matrix instead. If volume and polarity do not dominate, this challenges their assertion and needs to be explained.

Thompson et al (Thompson et al, 1997) do not use an SP score; they prefer instead a vectorial measure. Their program CLUSTALX, a graphical user interface to the CLUSTALW multiple alignment package (Higgins et al, 1996), plots a graph of positional conservation beneath a visual display of a multiple alignment. In the column at position x in the alignment, Thompson et al consider the residue of the ith sequence, $s_i(x)$, to be a point $\mathbf{X}_i$ in K-dimensional space:

$$\mathbf{X}_i = \begin{pmatrix} M(a_1, s_i(x)) \\ M(a_2, s_i(x)) \\ \vdots \\ M(a_K, s_i(x)) \end{pmatrix},$$

where a_n is the nth symbol in an alphabet of K possible amino acids and $M(a,b)$ is similarity as judged by a substitution matrix. The consensus amino acid, which for columns that are not strictly conserved will be a hypothetical construct, is the centre of gravity of all points from the column, $\overline{\mathbf{X}}$, ie,

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_i^N \mathbf{X}_i.$$

The degree of conservation among these points is then related to the average Euclidean distance of all points

[4] But their score is included in this section because the effect is much the same.

from the consensus point:

$$C_{Thompson} = p_{amino} \times \frac{1}{N}\sum_{i}^{N}\left|\overline{\mathbf{X}} - \mathbf{X}_i\right| , \tag{2.9}$$

where p_{amino} is the fraction of symbols that are not gaps.

All three of C_{Karlin}, V_{Armon} and $C_{Thompson}$ correctly order columns (a) to (f) and (g) to (i) in Figure 2.3 and have mathematically continuous output spaces. $C_{Thompson}$ has the aesthetic advantage of defining a consensus point in amino acid space. Although this may not correspond to a particular amino acid, the closest amino acid could easily be found.

Pilpel & Lancet (Pilpel & Lancet, 1999) use a mutation data score to help analyse amino acid variability in olfactory receptor sequences. They define their score as

$$V_{Lancet} = \sum_{a}^{K}\sum_{b}^{K}\frac{p_a p_b}{M(a,b)} ,$$

where p_a is the fractional frequency of amino acid a in the aligned column, the alphabet of amino acids is $K = 20$ and $M(a,b)$ is a BLOSUM62 or similar substitution matrix. Although this score is not directly comparable to those above, the following alteration makes it extremely similar to the tug-of-war definition of C_{Karlin} (equation 2.6):

$$C_{NotLancet} = \sum_{a}^{K}\sum_{b}^{K} p_a p_b M(a,b) .$$

This intermediate score now has properties almost identical to that of C_{Karlin}, so further discussion is best focused on how $C_{NotLancet}$ differs from V_{Lancet}, ie, the placing of the term $M(a,b)$. Having $M(a,b)$ as a denominator blights V_{Lancet} with an idiosyncratic output space. For instance, if a column contains a and b such that $M(a,b) = 0$, then V_{Lancet} for this column will be infinity. This will certainly happen if M corresponds to a raw BLOSUM62 matrix (Figure 2.6). But it is also difficult to imagine a reasonable matrix normalization that would avoid this problem. Thus, V_{Lancet} fails on at least two counts: its mathematical properties make it awkward to use and it fails to account for gaps.

2.2.5 Stereochemically sensitive entropy scores

The entropy scores discussed in section 2.2.2 quantified symbol diversity in an elegant and intuitive way. Their problem was they failed to account for stereochemistry. Scores in this section represent attempts to build stereochemical sensitivity into the entropy model.

Entropy measures the diversity of N symbols from an alphabet κ comprising K types. The difference between a symbol of one type and that of another is ineluctably uniform. What can be changed is how κ partitions amino acid space. Recognizing the deficiencies of symbol entropy scores, Mirny & Shakhnovich (Mirny & Shakhnovich, 1999) use the following stereochemically sensitive entropy score to analyse conservation at protein structure cores:

$$V_{Mirny} = \sum_{i}^{K} p_i \ln p_i ,$$

ie, Shannon's entropy, where $K = 6$ and κ is the set (eligible amino acids in square brackets): aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE] and special conformations [GP].

Williamson (Williamson, 1995) provides a similar score, which he uses to look at sequence variability

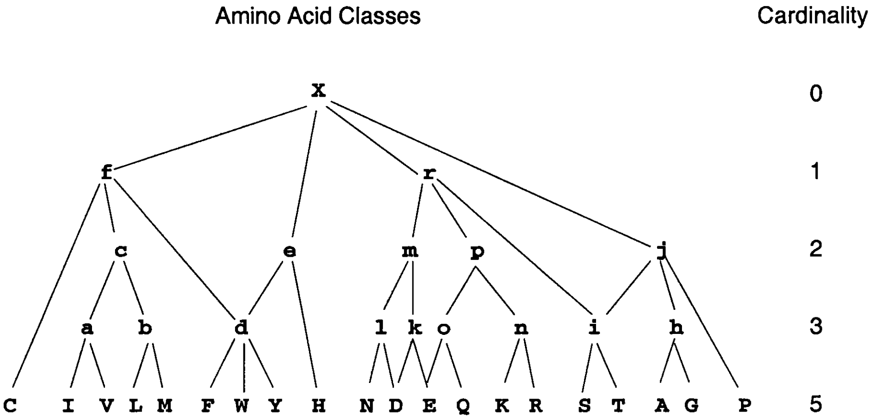


Figure 2.7: Amino acid class hierarchy used in PIMA.
Upper case characters are amino acids, lower case characters are amino acid classes. X is a wild-card character of any type, including a gap. In its original use, which was pairwise alignment, the match score between two aligned amino acids is the cardinality of the smallest class that includes both elements. This use has been extended by $V_{Goldstein}$ (see text). Adapted from ref (Smith & Smith, 1992).

in transporter proteins:

$$V_{Williamson} = \sum_{i}^{K} p_i \ln\left(\frac{p_i}{\overline{p}_i}\right),$$

where $\overline{p}_i$ is the fractional frequency of type i in the whole alignment, $K = 9$ and κ is the set: [VLIM], [FWY], [ST], [NQ], [HKR], [DE], [AG], [P] and [C]. An improvement over the scores discussed in section 2.2.2, V_{Mirny} and $V_{Williamson}$ correctly order columns (a), (b) and (c) from Figure 2.3 as more conserved than columns (d), (e) and (f). They also order columns (g) to (i) correctly. However, unlike the scores in section 2.2.2, neither can distinguish among (a), (b) and (c) or among (d), (e) and (f). So grouping residues in this way has its price. Moreover, neither score accounts for gaps. In their analysis, Mirny & Shakhnovich acquit themselves of this charge by choosing to ignore columns that contain gaps. But the problem of how to model gaps in the entropy score (see section 2.2.2) remains. One solution, which has not been implemented, might be to factor in gaps at the end using a scalar such as in equation 2.9.

Incorporating stereochemistry into an entropy score involves compromise. But does the choice have to be so stark: between, on the one hand, a robust but stereochemically insensitive description of relative amino acid frequencies (eg, $V_{Schneider}$); on the other, a clumsy partitioning of the 20 amino acids that accounts for some stereochemistry but ignores relative frequencies within a partition (eg, V_{Mirny})? There is a third way. In their pattern-induced multi-sequence alignment (PIMA) algorithm, Smith & Smith (Smith & Smith, 1992) use a hierarchical clustering of amino acids to extract sequence profiles from multiple alignments (Figure 2.7). Given the set amino acid types from an aligned column, PIMA finds the smallest possible "covering class" in the hierarchy that includes them all. For instance, the amino acids F, W and H are subsumed by the covering class [FWYH] (superset "e" in Figure 2.7). A conservation score has been suggested (although not implemented) that uses Shannon's entropy to assess the diversity of symbols in a column, then factors in the exclusivity of the smallest subset to which those symbols belong (Richard Goldstein, personal communication), eg,

$$V_{Goldstein} = f\left(\sum_{i=1}^{K} p_i \ln p_i, \gamma\right),$$

where $K = 21$ (ie, 20 amino acids plus one gap symbol), γ is the cardinality of the smallest covering class (see Figure 2.7) and f is some combining function. Gaps are penalized because they belong to only the largest superset and so have low cardinality ($\gamma = 0$). This is much like a synthesis of V_{Taylor} and $V_{Schneider}$. But because the PIMA hierarchy is ad hoc, and f is likely to be so, any statistical rigour potentially conferred by the use of entropy is lost.

2.2.6 Weighted scores

Scores in this category attempt to normalize against sequence redundancy in the alignment.

2.2.6.1 Background

Normalizing against redundancy is a concern not only for scoring conservation but also for building sequence profiles. Sequence weighting has thus received attention from exponents of both fields and there are a large number of methods to choose from. A selection of these methods is reviewed here.

The weight of a sequence is inversely related to its genetic distance from other sequences in the alignment. The simplest formulation is that given by Vingron & Argos (Vingron & Argos, 1989), where the weight of a sequence is equal to its average distance from all other sequences, ie,

$$w_i = \frac{1}{N-1} \sum_{j \neq i}^{N} d\left(s_i, s_j\right) , \tag{2.10}$$

where w_i is the weight of the ith sequence, s_i, and $d\left(s_i, s_j\right)$ is the genetic distance between the ith and jth sequences, measured as their percentage identity or some more sophisticated measure. Sander & Schneider incorporate a variation of this into their HSSP database (Sander & Schneider, 1991). They define the weight of a sequence in terms not only of sequence distance but also of the weights of all other sequences:

$$\lambda w_i = \sum_{j \neq i}^{N} w_j d\left(s_i, s_j\right) ,$$

where λ is a scaling constant. Expressed in the above form, this apparently circular definition can be solved as an eigenvalue problem. This self-consistency is aesthetically appealing but, because it makes the weight calculation more complex and since Sander & Schneider do not justify it, an unnecessary mathematical flourish.

Another formulation attempts to maximize the spread of data in aligned columns using a metric related to symbol entropy (Henikoff & Henikoff, 1994). This method first weights sequences at individual positions in an alignment, then combines position weights to give sequence weights. The weight of the ith sequence at position x is

$$w_{i_x} = \frac{1}{k_x n_{x_i}} ,$$

where k_x is the number of amino acid types present in column x and n_{x_i} is the frequency of the ith sequence's amino acid at that position. For example, position (b) in Figure 2.3 contains two amino acid types, D and E. The entropy at this position would be maximal if these two types were evenly distributed, ie, if half the column was D and the other half E. To accomplish this, one can weight sequences $\{1..10\}$ such that the proportions of D and E are equal, ie, let sequence 10 have weight $1/(2 \times 1) = 0.5$ and let sequences $\{1..9\}$ each has weight $1/(2 \times 9) = 0.056$. Averaging along all positions in an alignment, each sequence then has

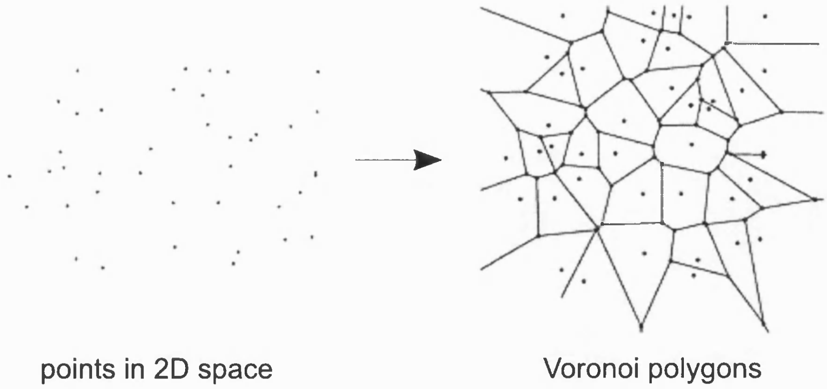


Figure 2.8: Voronoi diagram
Neighbouring points in a two-dimensional space are separated by a network of planes. Each plane is defined by the perpendicular to the bisector of two neighbouring points. Sibbald & Argos liken each sequence to a point and the volume of the surrounding Voronoi polygon to the weight of that sequence.

weight

$$w_i = \frac{1}{L}\sum_{x}^{L} w_{i_x},$$

where L is the length of the alignment.

Sibbald & Argos apply Voronoi diagrams to the problem of sequence weighting. A Voronoi diagram is a geometric structure that divides space around a set of points or objects. Given a set of points in two-dimensional space, the perpendicular bisector between each pair of neighbouring points is calculated. These bisectors are then extended until they all join up, forming polygons around the points as in Figure 2.8. Analogously, Sibbald & Argos consider sequences in an alignment to be a cloud of points in high dimensional space. They then apply the Voronoi procedure, defining polyhedra around each point, and take the weight of a sequence as the volume of its surrounding polyhedron. The more isolated a sequence is, the larger its polyhedron and the greater its weight. Sibbald & Argos estimate volumes of the polyhedra by filling the high dimensional space with random sequences. They show their method calculates more intuitive weights than the earlier method of Vingron & Argos in equation 2.10. However, the margin of difference is small and the Voronoi method is inconsistent, producing inexplicably different weights for equally redundant sequences.

Weighting methods popular in the construction of sequence profiles tend to rely on a phylogenetic tree of the multiple alignment. Thompson et al (Thompson et al, 1994) propose a weighting scheme based on Kirchhoff's laws, which describe how charge and voltage are distributed in an electrical circuit. They view the tree as a system of wires and nodes, and apply a voltage to the root. Kirchhoff's Current Law enforces conservation of charge: moving from the root to the leaves, it distributes current at each node so that the amount of charge entering at the root equals the amount exiting from the leaves. The current exiting at a leaf is taken as the weight of the corresponding sequence. So far so good. But the current entering a node is not necessarily distributed equitably among the outputs (branches). Rather, this distribution is governed by Kirchhoff's Voltage Law, which apportions greater current to the branch with more leaves. This inequitable distribution may be sensible for electric currents or water systems but it acts contrary to the motives of sequence weighting. Given a node that bifurcates into a highly populated subfamily and a sparsely populated one, the highly populated subfamily will receive the larger share of current and thus be upweighted.

Gerstein et al (Gerstein et al, 1994) approach the tree from the opposite direction: they start with the leaves and work up to the root, each sequence accumulating a share of the branch length as its weight. More sophisticated methods include those of Altschul et al (Altschul et al, 1989) and Eddy et al (Eddy et al, 1995) and are discussed at length in Durbin et al (Durbin et al, 1998). Tree-based weighting schemes are more assumptive than those based only on the alignment. After all, many plausible trees can describe a single alignment. Choosing one, even if it is the most probable, introduces additional uncertainty and thus hidden complexity.

2.2.6.2 Scores

Sequence weighting can be more easily incorporated into some scoring models than others. The sum-of-pairs model (page 43) accumulates contributions on a per sequence-pair basis. This provides an obvious placing for sequence weights. For entropy scores, which bundle amino acids according to type and disregard which sequence each came from, the placing is less obvious. Perhaps for this reason, weighted scores have tended to follow the SP model.

Landgraf et al (Landgraf et al, 1999) use the following score to extend the evolutionary trace method of Lichtarge et al (Lichtarge et al, 1996),

$$V_{Landgraf}(x) = \frac{1}{N}\sum_{i}^{N}\sum_{j>i}^{N}\left(w_i D\left(s_i(x), s_j(x)\right) + w_j D\left(s_j(x), s_i(x)\right)\right), \tag{2.11}$$

where $s_i(x)$ is the amino acid at position x of the ith sequence and w_i is the weight of sequence s_i as calculated by the Voronoi scheme of Sibbald & Argos (section 2.2.6.1). $D(a,b)$ measures the dissimilarity of the amino acids a and b and is calculated as

$$D(a,b) = \frac{m(a,a) - m(a,b)}{m(a,a)},$$

where m is the Gonnet substitution matrix (Benner et al, 1994). One of the first things to notice about D is its asymmetry. Intuitively, the difference between two amino acids is commutative such that $D(a,b) = D(b,a)$. However, because, as in most matrices, the diagonal scores in the Gonnet matrix differ depending on the amino acid, there are many cases when $m(a,a) \neq m(b,b)$ and therefore $D(a,b) \neq D(b,a)$. Landgraf et al recognize this inconsistency and hedge their bets in equation 2.11, with a sum of the form $w_i D(a,b) + w_j D(b,a)$. However, because this may give a different result from $w_j D(a,b) + w_i D(b,a)$, their handling of D's asymmetry is somewhat arbitrary.

Sander & Schneider do not use sequence weights as such (Sander & Schneider, 1991). Rather, they modify pairwise comparisons by the genetic distance between the sequences being compared:

$$C_{Sander}(x) = \lambda\sum_{i}^{N}\sum_{j>i}^{N} d\left(s_i, s_j\right) m\left(s_i(x), s_j(x)\right),$$

where $d(s_i, s_j)$ is the distance between sequence s_i and s_j measured as 100% minus their percentage identity in the alignment, m is the Dayhoff substitution matrix (Dayhoff et al, 1978), and λ scales C_{Sander} to range $[0,1]$, ie,

$$\lambda = \left(\sum_{i}^{N}\sum_{j>i}^{N} d(s_i, s_j)\right)^{-1}.$$

Sequence	Column x
s_1	W
s_2	Q
s_{3_1}	R
$\vdots$	$\vdots$
s_{3_n}	R

Figure 2.9: Column from a redundant sequence alignment.
Sequence s_{3_i} is the ith copy of sequence s_3. See text for details.

There are three obvious ways an SP score can incorporate a notion of sequence weighting. C_{Sander} uses one, $V_{Landgraf}$ uses another. The remainder of this section will show the choice of how to incorporate weighting into the score may be more important than the choice of weighting metric. It will also show the strategies of C_{Sander} and $V_{Landgraf}$ are surprisingly inferior to a strategy that looks only slightly different.

Consider an alignment of three sequences that are all equally different from one another. The column at position x in the alignment contains three different amino acids. The first sequence, s_1, has a W at this position, sequence s_2 has a Q and s_3 an R. Because the sequences are uniformly different, this alignment is "ideal" and requires no sequence weighting. Applying a simple unweighted sum-of-pairs score gives the result

$$C_{Simple}(x_{ideal}) = \sum_{i}^{N} \sum_{j>i}^{N} M(s_i(x), s_j(x)) = M(\mathtt{W},\mathtt{Q}) + M(\mathtt{W},\mathtt{R}) + M(\mathtt{Q},\mathtt{R}), \qquad (2.12)$$

where M is a symmetric similarity measure. Now add duplicates of sequence s_3 to the alignment to make it redundant and in need of sequence weighting. Figure 2.9 shows column x, which now contains one W, one Q and n Rs, corresponding to the n duplicates of s_3. Applying C_{Simple} to the new alignment gives the result

$$C_{Simple}(x) = M(\mathtt{W},\mathtt{Q}) + n\,(M(\mathtt{W},\mathtt{R}) + M(\mathtt{Q},\mathtt{R})) + \frac{n(n-1)}{2} M(\mathtt{R},\mathtt{R}). \qquad (2.13)$$

Clearly, as n increases two undesirable things happen. First, the $M(\mathtt{W},\mathtt{Q})$ term vanishes out of existence. Second, the spurious $M(\mathtt{R},\mathtt{R})$ term dominates. A good weighted SP score applied to the redundant alignment should at best reproduce the result in equation 2.12, at least moderate the affects of increasing n, and at worst reproduce the result in equation 2.13.

Let the distance between sequences $d(s_i, s_j)$ be 0 if $s_i = s_j$ and 1 otherwise. A C_{Sander}-like modification to C_{Simple} gives

$$C_{Distance}(x) = \sum_{i}^{N} \sum_{j<i}^{N} d(s_i, s_j) M(s_i(x), s_j(x)),$$

which, when applied to the redundant position x gives

$$C_{distance}(x) = M(\mathtt{W},\mathtt{Q}) + n\,(M(\mathtt{W},\mathtt{R}) + M(\mathtt{Q},\mathtt{R})). \qquad (2.14)$$

This is certainly an improvement on C_{Simple} because $M(\mathtt{R},\mathtt{R})$ has been factored out. However, it still has the problem that as n increases, $M(\mathtt{W},\mathtt{Q})$ disappears and the only effective comparisons are those involving s_3. A $V_{Landgraf}$-like modification (ignoring inconsistencies) to C_{Simple} gives

$$C_{sum}(x) = \sum_{i}^{N} \sum_{j>i}^{N} (w_i + w_j) M(s_i(x), s_j(x)).$$

For simplicity, we calculate w_i similarly to Vingron & Argos (equation 2.10) as

$$w_i = \sum_{i}^{N} d(s_i, s_j) .$$

This is reasonable because comparisons of this method with other weighting methods by Gerstein et al and Sibbald & Argos showed it was only slightly inferior. According to this scheme, $w_1 = w_2 = n+1$, whereas the weight of a single duplicate of s_3 is $w_3 = 2$. Applying C_{sum} to the redundant column gives

$$C_{sum} = (2n+2)M(\mathtt{W},\mathtt{Q}) + (n^2+3n)\left(M(\mathtt{W},\mathtt{R}) + M(\mathtt{Q},\mathtt{R})\right) + (2n^2-2n)M(\mathtt{R},\mathtt{R}) . \qquad (2.15)$$

This result is better than 2.13 but worse than 2.14. $M(\mathtt{W},\mathtt{Q})$ will still disappear because it increases with linearly with n whereas the other terms increase geometrically. $M(\mathtt{R},\mathtt{R})$ is also present. A third simple strategy is

$$C_{product}(x) = \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j M(s_i(x), s_j(x)) ,$$

which, when applied to the column in Figure 2.9 gives

$$C_{product} = (n^2+2n+1)M(\mathtt{W},\mathtt{Q}) + (2n^2+2n)\left(M(\mathtt{W},\mathtt{R}) + M(\mathtt{Q},\mathtt{R})\right) + (2n^2-2n)M(\mathtt{R},\mathtt{R}) . \qquad (2.16)$$

All terms are now on an equal footing with respect to n. Result 2.16 is clearly better than result 2.15 or 2.13. It is arguably more desirable than result 2.14 in that, although the spurious $M(\mathtt{R},\mathtt{R})$ features, no term disappears with increasing n.

2.3 Score used herein: C_{Valdar}

In this work, we use a sequence-weighted sum-of-pairs score. It is defined as follows:

$$C_{Valdar}(x) = \lambda \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j M(s_i(x), s_j(x)) ,$$

where N is the number of sequences, $s_i(x)$ is the amino acid of ith sequence at position x in the alignment, w_i is the weight of sequence s_i, $M(a,b)$ is the similarity of the amino acids a and b, and λ scales C_{Valdar} so that it ranges between 0 (maximally variable) to 1 (maximally conserved), ie,

$$\lambda = \left(\sum_{i}^{N} \sum_{j>i}^{N} w_i w_j \right)^{-1} .$$

The weight of a sequence is calculated according the scheme of Vingron & Argos (Vingron & Argos, 1989):

$$w_i = \frac{1}{N-1} \sum_{j \neq i}^{N} d(s_i, s_j) .$$

In this equation $d(s_i, s_j)$ is the distance between sequences s_i and s_j, and is calculated as

$$d(s_i, s_j) = 1 - \frac{1}{n(Aligned_{ij})} \sum_{x \in Aligned_{ij}} M(s_i(x), s_j(x)) ,$$

where $Aligned_{ij}$ is the set of all positions that manifest an amino acid in one or both of s_i and s_j, and $n(Aligned_{ij})$ is the size of this set. The comparison matrix M is a linear transformation of the substitution matrix m such that M takes values in the range [0,1] and all exchanges involving a gap score 0, ie,

$$M(a,b) = \begin{cases} \frac{m(a,b)-\min(m)}{\max(m)-\min(m)} & \text{if } a \neq \text{gap and } b \neq \text{gap} \\ 0 & \text{otherwise} \end{cases} .$$

The matrix m is a modified version of the pairwise exchange table (PET) (Jones et al, 1992), itself an updated version of the Dayhoff matrix (Dayhoff et al, 1978). The modified PET differs from the original in that all diagonal elements are set to a constant score, that score being the rounded average of diagonal elements in the unmodified matrix.

C_{Valdar} correctly orders columns (a) to (f), (g) to (i), and (j) and (k) in Figure 2.3. It fulfils or partially fulfils all criteria laid out in section 2.1.2; its output space is continuous and bounded; it accounts for amino acid frequency, like other SP scores, using the tug-of-war scheme (ie, self-similarity vs substitution); it quantifies stereochemical diversity uncompromisingly with a full substitution matrix; gaps incur a constant penalty; and it uses sequence weighting to normalize against redundancy in the alignment. C_{Valdar} weights its sequences using one of the simplest schemes available. Although there are other schemes that give marginally superior results, the Vingron & Argos weighting is simpler, makes fewer assumptions and at least gives consistent answers. C_{Valdar} incorporates its sequence weights shrewdly. By using the multiplicative scheme of $C_{product}$ (see page 52), its scores resist the distorting effect of many duplicate sequences in the alignment.

C_{Valdar} is not as simple as, say, the entropy score V_{Mirny} or the SP score C_{Karlin}, but it trades simplicity for sophistication in an economical way. C_{Valdar} attempts far more than these measures, but beyond adjustments specifically pertaining to gaps, stereochemistry and weighting, it incorporates no additional variables or transformations.

The normalization of m, the substitution matrix of C_{Valdar} is crude. That of C_{Karlin} is mathematically more elegant. However, C_{Valdar}'s normalization is more respectful of off-diagonal scores. These precious substitution probabilities are untouched and only the diagonal scores, which are anyway antithetical to the scoring of conservation, are affected.

2.4 A generalized formula for scoring conservation

No score is perfect. But some scores are less perfect than others. Scores discussed later in the survey tended to satisfy more of the criteria outlined in sections 2.1.1 and 2.1.2 than those discussed earlier. Shannon's entropy offered an elegant way to measure diversity among uniformly different symbols, but faltered when accounting for stereochemistry. Property-based scores (section 2.2.3) respected stereochemistry but failed to register symbol diversity. The most successful compromises were seen in the sum-of-pairs scores, although they exposed some limitations of using substitution matrices. Sum-of-pairs scores also seemed to be the most amenable to sequence weighting, although the review above is unlikely to be comprehensive.

C_{Valdar} is a compromise. In opting for a sum-of-pairs architecture it trades the mathematical elegance of Shannon's entropy for the rich stereochemical sensitivity bestowed by a substitution matrix. Gaps are incorporated in an ad hoc fashion, grafted on to the matrix. Any clumsiness of C_{Valdar} can be justified post hoc: it works, giving results consistent with intuition.

So far this chapter has mainly discussed scores following either the entropy or the substitution matrix model. But is this dichotomy inevitable? One could devise a score that plays entropy and mutation data to their relative strengths by keeping the assessment of relative symbol frequencies and the assessment of stereochemistry separate. An example of such a score is considered here.

Positional variability may be seen to have three elements:

1. symbol diversity, normalized to take account of sequence redundancy;

2. stereochemical diversity;

3. gaps.

For a given position, each element can be assigned a score that measures the extent to which it describes that column. Let t be the normalized symbol diversity (diversi*t*y), let r be the stereochemical diversity (ste*r*eochemistry) and let g be the gap cost (*g*ap). For convenience, all measures are continuous and bounded in the range 0 to 1, where 0 means that element is not present and 1 means that element is at its maximum. For instance, $r = 0$ means there is no stereochemical diversity at the position whereas $r = 1$ means the position could not be any more stereochemically diverse. Conservation is a function of these three variables. More intuitively, an assessment of conservation can be seen as a three pronged attack: a position is criticized on its symbol diversity, its stereochemical diversity and its gappyness. For a position x, we can write

$$C_{trident}(x) = (1 - t(x))^{\alpha} (1 - r(x))^{\beta} (1 - g(x))^{\gamma} . \tag{2.17}$$

The exponents α, β and γ weight the importance of each element. For the moment, suppose they are all equal to one. If position x is strictly conserved, then $C_{trident} = (1-0) \times (1-0) \times (1-0) = 1$. As position x becomes more afflicted with gaps, stereochemical diversity or symbol diversity, $C_{trident}$ drops towards zero. The relative impacts of these three elements on the conservation score were rigidly prescribed in C_{Valdar}. In $C_{trident}$, however, the sharpness of each prong may be adjusted freely to suit the purpose of the user. For example, if $C_{trident}$ with $\alpha = \beta = \gamma = 1$ is too lenient on gaps and too strict on stereochemistry for a particular application, one could instead try $\alpha = 1$, $\beta = 1/2$ and $\gamma = 2$.

$C_{trident}$ is so far more a convenient division of labour than a score, since it is open how any particular prong is defined. To make the score more concrete, we can start by specifying t as Shannon's entropy:

$$t(x) = \lambda \sum_{a}^{K} p_a \log_2 p_a ,$$

where K, the alphabet size, is 21 (20 amino acids plus one gap symbol) and p_a is the probability of observing the ath symbol type. λ_t scales the entropy to range [0,1] and is defined as

$$\lambda = \log_2(\min(N, K)) ,$$

where N is the number of sequences in the alignment, so that $t(x)$ can reach its maximum of one even when there are fewer that K amino acids in the column. Sequence weighting can be incorporated into Shannon's entropy by normalizing each p_a thus

$$p_a = \sum_{i \in \{i : s_i(x) = a\}} w_i ,$$

where w_i is the weight of the ith sequence and $s_i(x)$ is the symbol type at position x in that sequence. In words, the probability of observing symbol type a is the summed weight of sequences manifesting a.

Ideally, the sum of all weights should be one. The most apposite weighting scheme, which is related to an entropy model, is therefore that of Henikoff & Henikoff (Henikoff & Henikoff, 1994):

$$w_i = \frac{1}{L}\sum_{x}^{L}\frac{1}{k_x n_{x_i}},$$

where L is the length of the alignment, k_x is the number of symbol types present at the xth position and n_{x_i} is the number of times the symbol type manifested by the ith sequence occurs at that position (see section 2.2.6.1 for a fuller explanation).

The second prong of $C_{trident}$ measures stereochemical diversity but does not need to take account of symbol frequency or gaps. One candidate for this is $V_{Zvelibil}$ described in section 2.2.3. The one employed here uses a substitution matrix and is related to the model used in $C_{Thompson}$ (page 46). Let amino acid a be represented by a point $\mathbf{X}_a$ in 20-dimensional space such that

$$\mathbf{X}_a = \begin{pmatrix} M(a,a_1) \\ M(a,a_2) \\ \vdots \\ M(a,a_{20}) \end{pmatrix},$$

where a_i is the ith amino acid type. For example, the position of Cys in this space is defined by its mutational proximity to all other amino acids. $M(a,b)$ is the similarity between amino acids a and b judged by a normalized substitution matrix. One consistent normalization would be that of Karlin & Brocchieri (Karlin & Brocchieri, 1996) (equation 2.5). For a position x, the consensus amino acid type is calculated as point $\overline{\mathbf{X}}(x)$:

$$\overline{\mathbf{X}}(x) = \frac{1}{k_x}\sum_{a}^{k_x}\mathbf{X}_a,$$

where k_x is the number of amino acid types present in the column. The stereochemical diversity may be calculated as the average distance of observed amino acids from the consensus point:

$$r(x) = \lambda_r \frac{1}{k_x}\sum_{a}^{k_x}\left|\overline{\mathbf{X}}(x) - \mathbf{X}_a\right|,$$

where the scalar $\lambda_r = \sqrt{20\,(\max(M) - \min(M))^2}$ ensures $r \leq 1$.

The third prong of $C_{trident}$, the gap cost, is more straightforward. The more gaps, the less selective pressure is assumed to have acted at the position. Thus, $g(x)$ can be defined simply as the fraction of symbols in column x that are gaps.

$C_{trident}$ is not so much one score but a framework in which many different conservation scores can be imitated. For instance, if $\alpha = 1$, $\beta = 0$ and $\gamma = 0$, $C_{trident}$ resembles $C_{Schneider}$ (minus the weighting). However, because $C_{trident}$ raises more questions than it answers, it is better deployed in the analysis of conservation scores than in scoring conservation as such. After all, such a versatile framework can imitate uninformative scores as well as useful ones. An intriguing question is whether $C_{trident}$ can imitate C_{Valdar}. To investigate this, scores from C_{Valdar} were compared with scores from $C_{trident}$ for different values of α, β and γ. Specifically, C_{Valdar} and $C_{trident}$ were used to score all positions in six multiple sequence alignments. The similarity of the two scores was measured as Pearson's correlation coefficient of the two outputs. This was done for one thousand different sets of α, β and γ. To avoid unnecessary confounding

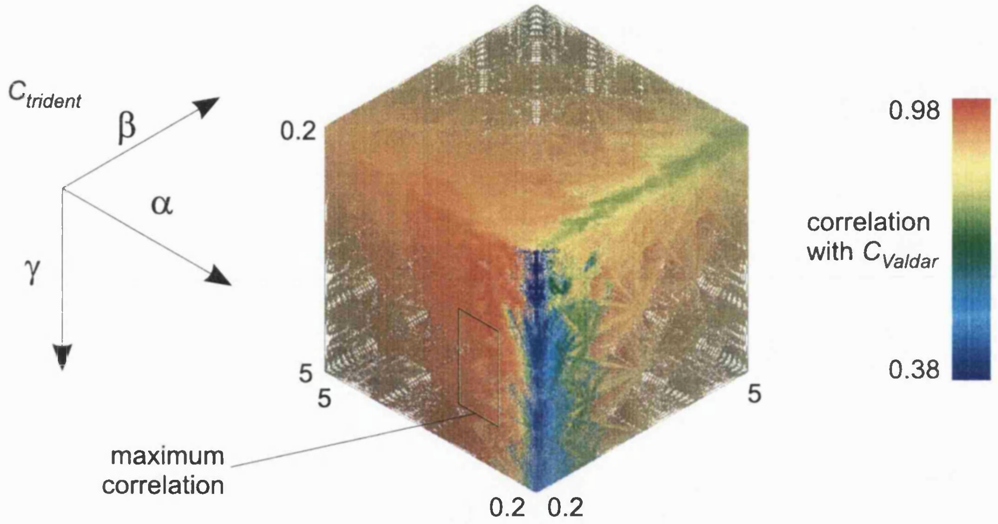


Figure 2.10: Similarity of C_{Valdar} to $C_{trident}$ under varying parameters α, β and γ. $C_{trident}$ is a flexible score for measuring residue conservation. It is parameterized by α, β and γ, which weight the relative importance of symbol diversity, stereochemistry and gap penalties respectively. Altering these parameters allows $C_{trident}$ to imitate a variety of inflexible conservation scores. The central cube represents the three-dimensional parameter space of $C_{trident}$. Colour is used to indicate the similarity, measured by the correlation coefficient of output (see text), of $C_{trident}$ to the concrete score C_{Valdar} at a particular point in this space (ie, for particular values of α, β and γ). Red areas indicate $C_{trident}$ is highly similar (has a correlation approaching 1) to C_{Valdar} at these values of α, β and γ. Blue areas indicate low similarity. The area of parameter space corresponding to the maximum correlation between the two scores (correlation coefficient=0.98 at $\alpha = 1$, $\beta = 0.5$, $\gamma = 3$) is approximately indicated on the diagram by an oblong box.

variables, the similarity matrix used for $C_{trident}$ was the same as that used for C_{Valdar}. The alignments used were those belonging to the six homodimer families discussed in chapter 3. This dataset comprised 1595 residue positions in all and contained data from 195 sequences. Figure 2.10 shows how the correlation between the two scores varied over the three-dimensional parameter space of α, β and γ. In this cursory investigation, the maximum correlation reached was 0.98 when $\alpha = 1$, $\beta = 0.5$ and $\gamma = 3$. This correlation seems high. However, because the dataset used is small and may not uniformly exercise all aspects of C_{Valdar}, this result should be considered only as a rough estimate. In particular, Figure 2.10 shows that when α and β are optimal, varying γ has little effect on the correlation. This reflects the small number of gaps in the six carefully compiled alignments. A gappier set of alignments might raise the profile of this third parameter. Acknowledging these caveats, it is nevertheless interesting to contrast the parameter set necessary to simulate C_{Valdar} with that required to simulate $C_{Shneider}$. For both concrete scores, $\alpha = 1$. This reflects the fact that C_{Valdar} and $C_{Shneider}$ both account for the relative frequencies of amino acids. The two scores differ on β. For C_{Valdar}, β is nonzero, indicating this score is sensitive towards stereochemistry whereas for $C_{Shneider}$, $\beta = 0$, indicating stereochemistry is ignored. C_{Valdar} penalizes gaps whereas $C_{Shneider}$ does not. Similarly, $\gamma = 3$ for C_{Valdar}, indicating this score's acknowledgement of gaps, whereas for $C_{Shneider}$, which does not penalize gaps, $\gamma = 0$.

2.5 Conclusions

This chapter has reviewed seventeen scores (not including C_{Valdar} or $C_{trident}$) and several distinct approaches for quantifying evolutionary conservation at an aligned position. No score achieved both biological and statistical rigour. The most meaningful scores were relatively ad hoc. However, given the success of probabilistic sequence profiles (Eddy, 1996) (Mott, 2000), which are a different but related emprise, it seems likely that a statistically robust score is possible.

$C_{trident}$ combines the strengths of previously disparate approaches. Although its flexibility undermines any authority it has as a concrete score (and for this reason it is not considered outside chapter 2), it does provides a framework for dissecting the character of other scores. This kind of meta-analysis is interesting from an abstract theoretical point of view. It may also be useful in a more practical sense. Given a dataset of multiple alignments with "correct" scores – these scores might be inferred from orthogonal information relating to the importance of particular residues in structure or function – the parameters of α, β and γ could be optimized so that $C_{trident}$ imitates these scores. At present, however, such datasets are not available and even somewhat difficult to conceive.

The score that satisfies the requirements of a conservation measure better than any other surveyed here is C_{Valdar}. C_{Valdar} will therefore be used to quantify residue conservation in the remainder of this work. Although more sophisticated scores could be conceived, C_{Valdar} accords with intuition and it will be used to answer the fundamental questions addressed in the following chapters about the utility of conservation in protein-protein interface prediction. The extent to which C_{Valdar} achieves this aim will be judged in the Conclusions (Chapter 5).

Chapter 3

Analysis of conservation in homodimeric interfaces

3.1 Introduction

Being able to predict protein-protein interfaces is desirable. But to make such a prediction, one must first know something about protein-protein interfaces that distinguishes them from other parts of the protein surface. The commonality of physical and chemical properties among some interfaces has led at least two groups to propose prediction methods on a purely physical and chemical basis (Young et al, 1994) (Jones & Thornton, 1997b). However, it is likely that evolutionary information derived from the large number of available protein sequences could be at least as useful. After all, protein-protein interactions are not merely biophysical phenomena; they are phenotypes, ultimately answerable to natural selection.

This chapter investigates the extent to which inferred evolutionary conservation can guide the analysis of protein-protein interfaces. If, following the neutralist view of evolution, residues under functional constraints tend to resist substitution (because substitution is usually for the worse, according to that view) rather than embrace it, one would expect residues in a biologically important interface to be conserved. If this is so, then identifying an interface could be as straightforward as locating a conserved cluster of residues on the surface. In order to establish whether conservation can be used this way, the work presented here tests the following premise: that residues in interfaces are significantly more conserved than those on the rest of the surface.

Grishin & Phillips (Grishin & Phillips, 1994) tested a similar premise, that interface residues are significantly conserved with respect to all other residues in a protein, and concluded it to be false. They analysed five oligomeric enzymes. For each enzyme sequence, they identified which positions corresponded to residues in the structural core, the active site and the subunit interface. Then, for every pair of sequences in a multiple alignment of the oligomer, they compare the rate of evolution, which they define as the fractional sequence identity, at these positions with that over all positions in the protein. This comparison gives them a measure of how much slower mutations occur in active site, core or interface positions than on average over the whole sequence. They found active site residues were by far the most conserved, evolving 50 times slower than average, whereas core and interface residues were only slightly conserved, evolving 2 and 1.5 times slower respectively. Thus, although the interfaces were much less conserved than the active sites, they were still more conserved than the surface.

Grishin & Phillips's definition of conservation precludes substitutions of any kind. But such a strict definition misses more subtle patterns of conservation: those in which substitutions conserve physico-chemical characteristics. Complete invariance at active sites positions is common because these motifs frequently rely on precise arrangements of specific amino acids. In contrast, residues at the structural core do tolerate mutations but within only a limited range (Branden & Tooze, 1998). A more sensitive measure of the rate of evolution, one that accounts for conservative substitutions, might have brought the scores for active site, core and, possibly, interface positions closer together.

Herein, we investigate whether oligomer interfaces are significantly conserved with respect to the protein surface by studying in depth a small but strictly defined dataset of six homodimer families, each of which form two-chain complexes. To address this problem meaningfully, we determine the probability that a randomly chosen group of residues from the protein surface will be more or less conserved than the interface group. We estimate this probability for all six oligomer complexes in our dataset by simulation, performing a large number of trials to obtain the fraction of random selections that equal or better interface conservation. The trials are performed in two ways: "picking", in which groups of residues are chosen entirely at random from the surface, and "walking", in which randomly chosen groups may contain only residues that are structurally contiguous.

3.2 Materials and Methods

The following protocol was followed to test whether the interface residues of a component chain in a protein-protein complex are significantly conserved with respect to all residues on the surface of that chain. First, functionally equivalent homologues of the protomer are identified and aligned multiply. Second, each position in the protomer is given a score that measures the degree to which it is conserved in evolution as inferred from the multiple alignment. Third, each residue in the protomer is classified according to the extent it lies in the surface and the extent it participates in the interface. Last, the average conservation score for residues in the interface is compared with the distribution of average conservation scores for the same number of surface residues in randomly selected groups. This comparison allows us to estimate the probability that a randomly selected group will have an equal or better average conservation than the interface, and hence assess whether the conservation of an interface is statistically significant.

To put this work in the context of previous analyses of protein-protein interfaces, the interface conservation of each protomer is also examined using "surface patches", after Jones & Thornton (Jones & Thornton, 1997a).

3.2.1 Criteria for dataset

Component chains from oligomer complexes were chosen to fulfill the following criteria. The protomer to be studied must form a stable, symmetric complex with one other protomer to which it is identical or nearly identical such that the oligomer is homodimeric and the conservation of only one chain need be considered. The complex must be shown by its associated literature to be essential to the stability and correct function of the protein. The full wild-type complex must be available as a structure determined by X-ray crystallography in either the Protein Data Bank (PDB) (Bernstein et al, 1977) (Berman et al, 2000) or its derivative, the Protein Quaternary Structure File Server (Henrick & Thornton, 1998) (PQS; http://pqs.ebi.ac.uk/). Of all structures available for the complex, the structure chosen must have the best combination of the following properties: high resolution, inclusion of any bound cofactors that occur naturally; and, if applicable, the inclusion of a ligand similar in size and shape to that of the natural substrate. To enable the robust

Family					Representative Protomer Structure					
Name	Abbreviation	Description	References	Fold (Class/ Architecture/ Topology)	Name	Resolution (in Ångstroms)	Heteroatom / ligand groups	Data bank source	Organism

Alkaline phosphatase	AP	Widely distributed non-specific phosphomonoesterase.	DuBose & Hartl, 1990;
			Hulett et al, 1991;
			Kim & Wyckoff, 1991;
			Knowles, 1991
				Mainly beta/ sandwich/ immunoglobulin-like	1alk chain A	2	1xMg
							2xZn
							1xPO4
								PDB	Escherichia coli

Enolase	Enolase	Glycolytic enzyme that catalyzes the dehydration of 2-phospho-D-glycerate (PGP) to phosphoenolpyruvate (PEP).	Babbitt & Gerlt, 1997;
			Babbitt et al, 1996;
			Larsen et al, 1996;
			Zhang et al, 1997
				domain 1: alpha beta/ 2-layer sandwich/ enolase-like;
				domain 2: alpha beta/ barrel/ TIM barrel
					1one chain A	1.8	1xPEP
							1xMg
								PDB	Saccharomyces cerevisiae (baker's yeast)

Glutathione S-transferase	GST	Catalyzes conjugation of glutathione to a variety of electrophilic substrates (including carcinogens and anti-cancer drugs) and makes the latter easier for the host to metabolise.	Board et al, 1995;
			Board et al, 1997;
			Neuefeind et al, 1997;
			Rossjohn et al, 1997
				domain 1: alpha beta/ 3-layer(aba) sandwich/ glutaredoxin;
				domain 2: mainly alpha/ non-bundle/ glutathione S-transferase (subunit A, domain 2)
					1glq chain A	1.8	1xGTB (s-(p-nitrobenzyl) glutathione)	PDB	Mus musculus (house mouse) liver

Copper, zinc superoxide dismutase	SOD	Neutralizes superoxide radicals. Consistent homodimers only in cytoplasmic eukaryotic proteins.	Banci et al, 1998;
			Bordo et al, 1994, 1999;
			Getzoff et al, 1989
				mainly beta/ sandwich/ immunoglobulin-like	1xso chain A	1.49	1xCu
							1xZn
								PDB	Xenopus laevis (African clawed frog)

Streptomyces subtilisin inhibitor	SSI	Serine proteinase inhibitor that inhibits subtilisin strongly and other proteinases, including trypsin and chymotrypsin, to a lesser extent. Protomers inhibit one proteinase each to form E2I2 complex.	Hirono et al, 1984;
			Kojima, et al, 1993;
			Laskowski & Kato, 1980;
			Taguchi et al, 1997
				alpha beta/ 2-layer sandwich/ subtilisin inhibitor	2sic chain I	1.8	1xStreptomyces subtilisin (2sic chain E)	PQS	bacillus amyloliquefaciens

Triose phosphate isomerase	TIM	Catalyzes interconversion of D-glyceraldehyde 3-phosphate and dihydroxy acetone phosphate.	Borchert et al, 1994;
			Garza-Ramos et al, 1998;
			Gopal et al, 1999;
			Williams et al, 1999
				alpha beta/ barrel/ TIM barrel	1tph chain 1	1.8	2-phosphoglycolohydroxamate	PDB	gallus gallus (chicken) muscle

Table 3.1: Family information for the six homodimer families.

identification of a diverse set of homologues, the protomer should be represented in the CATH classification (Orengo et al, 1997) and according to that classification have only one structural domain. The protomer sequence must have non-fragment homologues in the SWISS-PROT protein sequence database (Bairoch & Apweiler, 2000) that are numerous (>10) and diverse (<70% mean pairwise sequence identity), and, by their annotation, share its function and multimeric state (see also Identification and alignment of homologues). Applying these criteria gave rise to six homodimer families (see Table 3.1).

3.2.2 Identification and alignment of homologues

Homologues for a given protomer are identified and aligned in two stages. At both stages, a homologue is included only if its annotation and associated references show unambiguously that it shares the protomer's function and precise multimeric state, and that it is a wild type protein and not a fragment. First, the sequence family of the protomer is identified in the CATH classification (Orengo et al, 1997). If there are at least three suitable representatives of that family, a multiple structural alignment of the protomer and these representatives is built using the CORA suite of programs (Orengo, 1999). Otherwise, a multiple sequence alignment from the ALIGN resource associated with the PRINTS database (Attwood et al, 1999) or one from the Pfam database (Bateman et al, 2000) is edited and used. This seed alignment is used to build a profile hidden Markov model (Eddy, 1996, and references therein), which in turn is used both to find more homologues in the SWISS-PROT sequence database and to align multiply the final set of homologues to the protomer sequence. This final alignment, referred to here as the "full alignment" for the family, is edited for high redundancy by removing the less well characterized or shorter of any two sequences that are more than 90% identical. Searching for homologous structures in CATH was performed with reference to the CATH Dictionary of Homologous Superfamilies (Bray et al, 2000). The construction of the profile and its application in searching for and aligning homologues were both performed with the HMMER2 software package (see http://hmmer.wustl.edu/). The filtering of large numbers of sequences by annotation was performed using the Sequence Retrieval System (Etzold et al, 1996). The removal of redundancy in the alignment was performed with the help of JalView (Clamp et al, 1998).

3.2.3 Scoring residue conservation

Each residue in the protomer of interest is assigned a numerical value *Cons* (ranging from 0 to 1) corresponding to the conservation of residue similarities at its position in the multiple sequence alignment. A value of 0 indicates the position is not conserved; a value of 1 indicates it is highly conserved. *Cons* corresponds the C_{Valdar} conservation score defined in section 2.3.

Here the terms "conservation score" and "residue conservation" will be used to denote either the value returned by *Cons* for a given residue, or, when applied to a set of residues, the average value of *Cons* for all residues in that set. These definitions are vital because they underpin the whole analysis of conservation.

3.2.4 Interface definition

Each residue in a protomer is assigned to one of the following disjoint sets: *Core*, *Exposed*, *Partially Buried*, or *Buried*. Qualitatively, *Core* residues are those in the structural core of the protein, *Exposed* residues are on the surface but do not participate in an interface, and *Partially Buried* and *Buried* describe residues that are on the surface of the protomer and participate in a multimer interface. Two further sets are referred to here: the *Surface* set, which is the union of the *Exposed*, *Partially Buried* and *Buried* sets, and contains all residues on the surface of the protomer; and the *Interface* set, which is used as a generic term for either the *Buried* set or the union of the *Buried* and *Partially Buried* sets. For clarity, *Buried* residues are said to form the "central zone" of the interface whereas *Partially Buried* residues form its "outer zone". Together, these sets define the "total interface".

The classes are assigned on the basis of solvent accessibility, which is calculated using NACCESS (Hubbard & Thornton, 1993), an implementation of the Lee & Richards (Lee & Richards, 1971) algorithm, with a probe sphere of radius 1.4 Å. A residue is deemed accessible if its relative accessible surface area (RSA) is > 5%, a cut-off devised and optimized by Miller et al (Miller et al, 1987). If a residue is accessible in the protomer it is in the *Surface* set, otherwise it is *Core*. If a residue in the *Surface* set loses RSA upon complexation it is in the *Interface* set, otherwise it is *Exposed*. If a residue in the *Interface* set is inaccessible (ie, < 5% RSA) in the multimer complex, it is in the *Buried* set, else it is *Partially Buried*.

3.2.5 Ligand-buried residues

Ligand-buried residues are defined here as residues that become inaccessible in the protomer upon inclusion of ligand groups. Together they comprise the ligand-buried site for a protomer. Because residues that participate in an active or allosteric site (referred to here generically as a "binding site") are typically both accessible and highly conserved, the inclusion of ligand-buried residues, which are usually a subset of the binding site residues, in the *Surface* set will clearly affect any calculation that compares conservation of interface and surface. To investigate the effect of conserved ligand-buried residues, all tests described below are carried out twice, once with these residues included, or "unmasked", and once with them excluded, or "masked".

3.2.6 Patch analysis of interface conservation

The conservation of the total interface of each unmasked protomer was examined using a variant of the "patch analysis" method of Jones & Thornton (Jones & Thornton, 1997a). In the original procedure, a set of roughly circular overlapping patches, each covering as many residues as the interface, is defined on the surface of the protomer. Quantitative properties of patches and their constituent residues can then be described in terms of their distributions over all patches and related to the extent those patches overlap with the interface. Herein, the average conservation of residues in a patch is the only property considered and this quantity is termed the "patch score".

3.2.7 Testing the significance of interface conservation

In order to assess the significance of conservation at a given interface the following null hypothesis, H_0, is tested: the average conservation of the *Interface* set is no higher than that obtained from an equal number of residues drawn without replacement from the *Surface* set by a random process. The negation of H_0 is the alternative hypothesis, H_1, which states that the *Interface* set has a higher average conservation than that of a set randomly selected in this way. A simulation experiment is performed to estimate the probability that H_0 is true. If the value of this probability (P value) falls below a certain threshold, customarily defined at 0.05, then H_0 is rejected in favour of H_1 and the conservation of the interface is considered statistically significant at the 5% level.

The P value expresses the probability that a selection of residues drawn from the surface by a random process will have an average conservation equal to or greater than that of the interface. This P value depends not only on the distribution of conservation over the surface and in the interface but also on the nature of the random process employed to make the selections. To ensure the test of H_0 is meaningful, it is a minimum requirement of any random process used that it is able to draw from the surface the set of residues corresponding to the interface. Two distinct random selection processes, "picking" and "walking", are employed here and these are described below.

For a given protomer surface, defined interface and random process, it is often computationally infeasible to enumerate all possible selections, compare the conservation of each to that of the interface, and hence evaluate p, the true P value. However, p can be estimated reliably enough by random sampling.

A trial is devised in which the random process selects *n(Interface)* residues from surface set and their average conservation is compared with that of the residues in the interface set. If t such trials are performed and each trial is independent of any other, then p can be estimated as $\widehat{p} = t_c/t$, where t_c is the number of trials in which the selection was at least as conserved as the interface. The greater the number of trials, more reliable the estimate, and when t is large, the expected accuracy of $\widehat{p}$ can be described formally by a confidence interval. A confidence interval is defined by a range, symmetric about the estimate and an associated probability that p, the true value, is contained somewhere within this range. The "99% confidence interval" for the unknown value of p is thus the margin of error expected for $\widehat{p}$ 99% of the time. This interval is given by $(\widehat{p}-2.58\sigma, \widehat{p}+2.58\sigma)$, where σ, the standard deviation of $\widehat{p}$, is equal to $\sqrt{\widehat{p}(1-\widehat{p})/t}$. To ensure a high degree of accuracy, the number of trials performed for a given a estimate is constant at 10 million, resulting in a margin of error of at most $\simeq$0.04% at least 99% of the time.

The P value for interface conservation is estimated stochastically as described above for each of the six protomers in the dataset. For each protomer, trials are performed under three variable conditions, giving rise to eight experiments per protomer. First, trials are performed using one of the two random selection processes, "picking" and "walking". Second, residues participating in an active or allosteric site are either included in the *Surface* set or masked out. Third, the *Interface* set is taken as either the *Buried* set (central zone) or the union $PartiallyBuried \cup Buried$ (total interface).

3.2.7.1 Picking: unconstrained selection of residues

"Picking" is the first of the two random processes used here for selecting a group of residues from the $Surface$ set of a protomer. For a given protomer, residues are drawn at random and without replacement from the $Surface$ set until the number drawn is equal to $n(Interface)$, the number of residues in the $Interface$ set. In picking, all selections occur with equal probability.

3.2.7.2 Walking: structurally constrained selection of residues

"Walking" is the second of the two random processes used here for selecting a group of residues from the $Surface$ set of a protomer. Walking selects groups of residues from the surface of a protomer by successively stepping from one residue to any residue in contact with it chosen at random. A walk starts at any residue chosen from the entire $Surface$ set. The walk is allowed to revisit residues any number of times, otherwise it could become trapped, but any particular residue is counted only once towards the final selection. The walk ends when the number of distinct residues visited is equal to the number of residues in the $Interface$ set. In this scheme, two residues, A and B, are considered "in contact" if the distance between the van der Waals spheres of at least one of A's atoms and at least one of B's atoms is no more than 1 Å. All walks are equiprobable but many walks may produce the same selection.

3.3 Results

We investigated interface conservation in six homodimer families (abbreviations in brackets): alkaline phosphatase (AP), enolase (Enolase), glutathione S-transferase (GST), copper-zinc superoxide dismutase (SOD), *Streptomyces* subtilisin inhibitor (SSI), and triose phosphate isomerase (TIM) (see Table 3.1).

Table 3.2 lists the number of residues in the central zone and total interface of each protomer representative, and Figure 3.1 shows graphically the residues that make up the total interface. The total interface for a family was typically contiguous and compact, though not particularly circular. Only the total interface of SSI was non-contiguous, in which one residue, Pro37, was separated from the closest of the others by 3 Å (Figure 3.1(SSI,a), (SSI,c)). An approximately linear relationship was observed both between the size of the central zone and total interface and between the size of the total interface and the number of surface residues. The largest total interface was that of AP and covered a third of that protomer's surface residues (Figure 3.1(AP,a) & Table 3.2). The smallest total interface, which covered barely a fifth of surface residues, belonged to GST (Figure 3.1(GST,a) & Table 3.2). The central zone was between 20% (GST) and $\simeq$40% (AP) of the size of the total interface.

Ligand-buried residues were identified in AP, Enolase, SSI and TIM (Table 3.3 & Figure 3.1). The ligand-buried sites of AP and Enolase both comprise two highly conserved residues situated in pockets near the interface (Figures 3.1(AP,a), (Enolase,a)).

SSI's ligand-buried site is on the opposite side of the protomer to that of the interface (Table 3.3 & Figure 3.1(SSI,b)) and consists of three residues, Met70, Cys71 and Pro72, which protrude into and block the active site of serine proteinase. These residues are centrally located within an inhibitory region SSI shares with other serine proteinase inhibitors (serpins). This region, the so-called "reactive site", stretches from Gly66 to Tyr75 and, between Met73 and Val74, contains the peptide bond used as bait for the catalytic triads of proteinase enzymes. In contrast with the other protomers, SSI's ligand-buried residues are not unanimously conserved (Laskowski & Kato, 1980) (Hill & Hastie, 1987) (Takeuchi et al, 1992) (Kojima et al, 1993). Phylogenetic analysis by Taguchi et al (Taguchi et al, 1997) suggests variability in this region may result from diversifying selection driven by the advantages of multi-specific inhibitors in the regulation of intrinsic proteases.

TIM's ligand-buried site was the largest of those studied and contains five highly conserved residues (Table 3.3 & Figures 3.1(TIM,a), (TIM,c)). TIM binds its substrate in a pocket created by the inside edge of its barrel topology. Although this pocket lies just outside the total interface, two ligand-buried residues, Asn11 and His95, belong to the central zone. This surprising observation results from the convoluted geometry of the interface in which the two component chains protrude so deeply into one another that one affects the solvent accessibility of residues that form the inside of the other's barrel.

Although GST and SOD are both enzymes and their representative structures included ligand groups (see Table 3.1 & Figures 3.1(GST,a), (SOD,a), (SOD,b)), no ligand-buried residues were detected in either. This reflects the strict definition of the ligand-buried site (see Methods) in which ligand-buried residues must lose all accessibility upon binding ligands. For example, Lys13 of TIM, which is known to play an important role in catalysis (Williams et al, 1999) and is highly conserved, touches TIM's substrate. However, because it is not completely buried by the substrate it does not qualify here as a ligand-buried residue.

Defining which amino acid types are conserved in interfaces is complex and beyond the scope of this paper. Residues in the representative protomer of AP (Figure 3.1(c)(d)) map directly to positions in AP's multiple alignment and so may host a number of amino acid types in varying proportions. Moreover, the

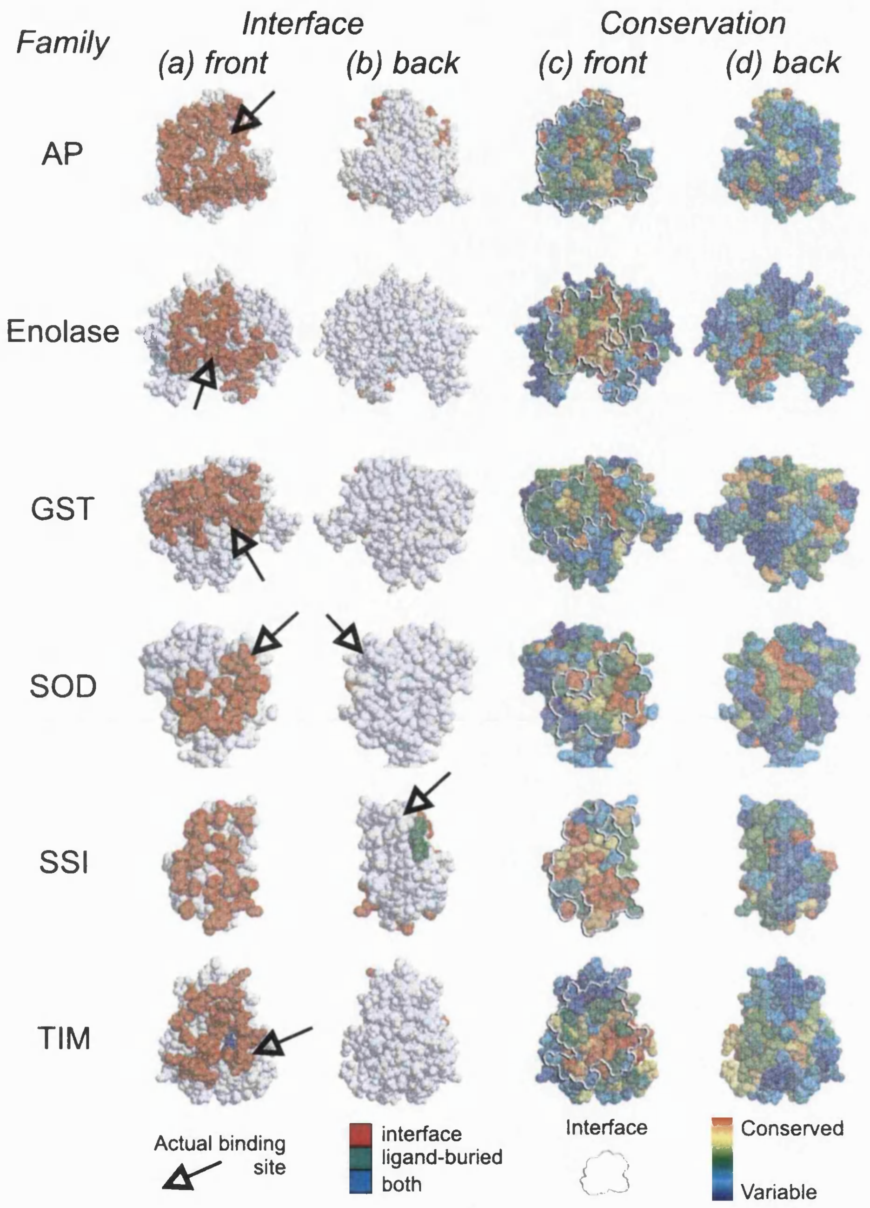


Figure 3.1: A table to show the location of interface and ligand-buried residues ((a),(b)), and residue conservation ((c),(d)) for six families of homodimers.
Protomer structures are elevated to show the interface head on in columns (a) and (c), and at a rotation of 180° about the y-axis in columns (b) and (d). In columns (a) and (b), ligand buried residues and residues belonging to the total interface are indicated (see Methods for how these classes are defined). Arrows indicate the approximate location of the actual binding site as it is defined in the literature. Ligand-buried residues, typically a subset of residues in the actual binding site, are detected in AP, Enolase, SSI and TIM. In the elevations presented here these residues are out of view for AP and Enolase, partially visible for TIM, and conspicuous in SSI. In columns (c) and (d), each residue is coloured by the rank of its conservation score among all other conservation scores in the protomer. Rank *Cons* is used instead of absolute *Cons* so that dispersion of conservation over the surface can be more easily visualized. A steel wire effect delineates the perimeter of the total interface. The table shows conservation is not distributed uniformly on the surface but in clusters, and that the interface, although it includes both highly and poorly conserved residues, is on average more conserved than not. Atom coordinates were obtained from the PDB (Bernstein et al, 1977) (Berman et al, 2000) and the PQS (Henrick & Thornton, 1998). Images were created using MOLSCRIPT (Kraulis, 1991) and Raster3D (Merritt & Bacon, 1997).

Family	Active site state	Number of surface residues	Mean surface *Cons*	S.D. surface *Cons*	Interface definition[1]	Number of interface residues	Mean interface *Cons*	Experiment type	P-value for interface[2]	Error (+/-)
AP	unmasked	289	0.59	0.21	central	37	0.71	picking	6.86E-5	6.76E-6
								walking	6.83E-3	6.72E-5
					total	96	0.63	picking	4.52E-3	5.47E-5
								walking	**7.02E-2**	2.08E-4
	masked	287	0.58	0.21	central	37	0.71	picking	3.87E-5	5.08E-6
								walking	3.12E-3	4.55E-5
					total	96	0.63	picking	2.36E-3	3.96E-5
								walking	5.00E-2	1.78E-4
Enolase	unmasked	263	0.72	0.22	central	20	0.89	picking	3.55E-5	4.86E-6
								walking	3.74E-2	1.55E-4
					total	52	0.82	picking	3.02E-5	4.48E-6
								walking	3.78E-2	1.56E-4
	masked	261	0.72	0.22	central	20	0.89	picking	2.57E-5	4.14E-6
								walking	3.18E-2	1.43E-4
					total	52	0.82	picking	2.18E-5	3.81E-6
								walking	3.13E-2	1.42E-4
GST	unmasked	158	0.45	0.16	central	6	0.71	picking	1.06E-4	8.40E-6
								walking	1.77E-3	3.43E-5
					total	30	0.52	picking	3.72E-3	4.96E-5
								walking	**7.09E-2**	2.09E-4
	masked	158	0.45	0.16	central	6	0.71	picking	1.10E-4	8.54E-6
								walking	1.74E-3	3.40E-5
					total	30	0.52	picking	3.72E-3	4.96E-5
								walking	**7.11E-2**	2.10E-4
SOD	unmasked	105	0.70	0.23	central	5	0.93	picking	8.73E-3	7.59E-5
								walking	2.75E-2	1.33E-4
					total	20	0.78	picking	3.60E-2	1.52E-4
								walking	**2.02E-1**	3.27E-4
	masked	105	0.70	0.23	central	5	0.93	picking	8.73E-3	7.59E-5
								walking	2.75E-2	1.34E-4
					total	20	0.78	picking	3.60E-2	1.52E-4
								walking	**2.02E-1**	3.27E-4
SSI	unmasked	91	0.71	0.23	central	7	0.85	picking	4.50E-2	1.69E-4
								walking	**1.59E-1**	2.98E-4
					total	27	0.84	picking	1.50E-4	9.99E-6
								walking	1.30E-2	9.25E-5
	masked	88	0.71	0.22	central	7	0.85	picking	4.53E-2	1.70E-4
								walking	**1.64E-1**	3.02E-4
					total	27	0.84	picking	1.43E-4	9.75E-6
								walking	1.35E-2	9.41E-5
TIM	unmasked	168	0.59	0.22	central	9	0.76	picking	1.30E-2	9.25E-5
								walking	**1.65E-1**	3.03E-4
					total	38	0.68	picking	1.92E-3	3.57E-5
								walking	**1.71E-1**	3.07E-4
	masked	163	0.58	0.21	central	8	0.73	picking	2.61E-2	1.30E-4
								walking	**1.63E-1**	3.02E-4
					total	36	0.66	picking	2.90E-3	4.39E-5
								walking	**1.43E-1**	2.86E-4

[1] central = "central zone", total = "total interface"

[2] $p \geq 0.05$ in bold

Table 3.2: P values and associated information calculated for the six homodimer families.

Family	Ligand-buried residues
AP	Thr102, Asp327
Enolase	Ser39, Lys345
GST	None
SOD	None
SSI	Met70, Cys71, Pro72
TIM	Asn11, His95, Glu65, Gly210, Gly232

Table 3.3: Ligand buried residues detected in the six homodimer families.

notion of conservation as a continuous quantity suggests no obvious cutoff at which "conserved" residues could be distinguished. However, for completely conserved positions, ie, those with a conservation score of 1, such an analysis is simple. By far the most common amino acid invariant at the interfaces of the six homodimers was glycine. This is probably because glycine does not have a side chain and so substituting it with an amino acid that does causes sterically unacceptable disruption of the interface. Arginine and valine were next most common but their numbers are low and so cannot be interpreted with confidence.

For each family, the statistical significance of residue conservation at the oligomeric interface was assessed by computing the probability (*P* value) that this conservation could have occurred by chance (see Methods). If the *P* value was less than the predefined cutoff 0.05, the associated interface was considered significantly conserved. *P* value calculations were performed under three variable conditions (described in Methods), giving rise to eight distinct *P* values per family. In addition to the significance tests, patch analysis was performed on each family whereby the conservation of residues at a homodimeric interface is compared with that of roughly circular overlapping patches defined on the surface of a constituent protomer.

3.3.1 Conservation in patches

The protomer dataset was analysed using surface patches. For each family, patches containing as many residues as the total interface were defined on the surface of the representative protomer and the average residue conservation of each patch, ie, its patch score, was calculated (see Methods). The number of patches defined, which related linearly to the size of the protomer, ranged from 87 in SSI to 235 in AP.

Figure 3.2 shows distributions of the patch scores for each family and reveals that, for all families, the average conservation of residues in the observed interface lies within the top quarter of the distribution. Specifically, the score of the interface coincides with the following percentiles: 77% (ie, lying just within the top 23% of the distribution) (TIM), 80% (SOD), 82% (GST), 84% (Enolase), 91% (AP), and 92% (SSI). It is more meaningful to compare the interfaces of different families based on relative patch rank in this way than by absolute conservation score because the latter depends on the extent and diversity of the underlying multiple sequence alignments.

The mean of a patch score distribution tends towards the mean conservation of surface residues in the corresponding protomer. The higher moments (eg, standard deviation, skewness and kurtosis) depend not only on the shape of the distribution of conservation scores for individual residues but also on the patch size and how conservation is dispersed about the surface. As expected, larger patches tend to give narrower distributions. For example, the variance of residue conservation for AP is similar to that of the other families (see Table 3.2) but, owing partly to the large number of residues in its total interface, its distribution of patch scores is markedly narrower (Figure 3.2(a)). The less uniformly the extremes of residue conservation are dispersed over the surface of a protomer, the greater the difference between the highest and lowest patch scores. Dispersion therefore affects not only the width of the distribution, causing it to be spread out if

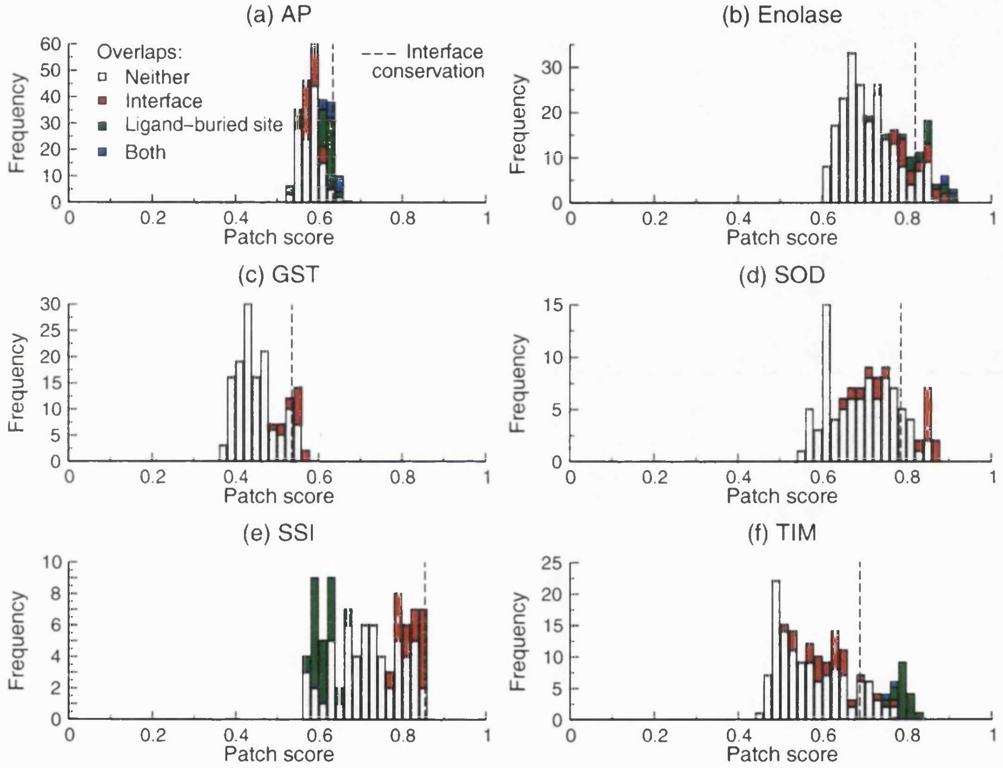


Figure 3.2: Distributions of patch scores for six families of homodimers.

The patch score, defined as the mean conservation score for all residues in a patch, is given in bins of width 0.02 along the x-axis. The number of patches that fall into a given patch score bin is presented on the y-axis. Stacking within histogram bars indicates the proportion of patches falling into a given bin that overlapped the interface, overlapped the ligand buried site, overlapped neither region and overlapped both (see Results for overlap criteria). A dashed line indicates the interface conservation, defined as the mean conservation score for all residues in the interface. The graphs show that patches overlapping the interface tend to score highly and that interface conservation consistently lies within the top quarter of the patch score distribution.

residues with high and low conservation cluster in space, but also the skewness and kurtosis. For example, if residues with high and low conservation cluster heavily at opposite sides of a protomer, most patches will contain many more residues from one extreme than from the other, with few patches straddling both poles equally to achieve the mean score. The resulting distribution will have a sunken appearance (negative kurtosis) such as that seen for SSI (Figure 3.2(e)). If the degree of concentration is greater at one pole than the other, the distribution will be correspondingly asymmetric (skewed) as seen for TIM (Figure 3.2(f)).

3.3.1.1 Overlap of the interface and ligand-buried site

Patches are deemed to overlap with the interface if they contain at least half of the interface residues, and overlap with the ligand-buried site if they contain all the ligand-buried residues. Overlap is defined differently to take account of the difference in size between these two regions and how much of each a patch can reasonably cover, ie, most patches that overlap some ligand-buried residues will overlap all of them whereas only one patch can cover all interface residues.

In fact, no patch overlapped any interface completely. The greatest percentage overlap achieved by a patch for a particular family ranged from 67% (AP and SSI) to 80% (SOD). Patches that overlapped the interface tended to be at least moderately and often highly conserved relative to other patches.

The stacked histogram for AP shows that patches that overlap with either the interface or the ligand-buried site occur throughout the distribution, but patches that overlap with both regions occur only among the higher ranks (Figure 3.2(a)). Patches that overlap the interface score variably despite the apparent high percentile ranking of AP's true interface because any one patch owes at least a third of its score to residues outside the interface. Patches that overlap both interface and ligand-buried site score highly because they include not only conserved interface and ligand-buried residues but also some of the conserved binding site residues that surround the ligand-buried site. The narrowness and symmetry of the AP distribution is, as mentioned above, partly explained by the large size of each patch but also reflects the unclustered dispersion of high and low conservation over the surface observed in Figures 3.1(AP,c), (AP,d).

The histogram for Enolase shows that overlaps with either the interface or the ligand-buried site occur almost exclusively at the top end of the distribution, with patches that overlap both taking the highest ranks (Figure 3.2(b)). Examining conservation at the surface of Enolase reveals a concentration of highly conserved residues around the ligand-buried site and in the region of the interface nearest to it (Figure 3.1(Enolase,c), (Enolase,d)). As for AP, optimally scoring patches tend to be those that cover both regions. The width and positive skewness of the Enolase distribution reflects the clustering of high conservation at the surface in the absence of any poorly conserved clusters.

Patches that overlap the interface are found at only the top end of GST's patch score distribution and all the highest scoring patches have interface overlap (Figure 3.2(c)). GST has smaller patches than Enolase, and, because the sequences that contribute to its alignment are more divergent, there is greater variance in the conservation of its surface residues (Table 3.2). Yet GST has the narrower distribution. This is because the dispersion of conserved residues over the surface of GST is far less clustered than for Enolase, so its patch scores tend to deviate less from the distribution mean (see Figures 3.1(GST,c), (GST,d)).

In SOD, overlap with the interface is split between the middle and top end of the distribution of patch scores (Figure 3.2(d)). This dichotomy results from a slight clustering of poorly conserved residues on one side of the interface along with a slight clustering of highly conserved residues on the other (Figures 3.1(SOD,c), (SOD,d)). Patches that overlap the side of the interface near the unconserved cluster have moderate conservation whereas those that overlap the other side have high conservation.

The histogram for SSI shows a striking separation of interface overlap, which is confined to the upper end of the distribution, from ligand-buried site overlap, which is found exclusively at the lower end. This separation arises because conservation over the surface of SSI is polarized, with a majority of highly conserved residues around the interface and a majority of poorly conserved residues around the hypervariable ligand-buried site.

In TIM, there is a smooth progression from those patches that overlap the interface, which are moderately conserved, to patches that overlap the interface and the ligand-buried site, which are well conserved, to patches that overlap the ligand-buried site, which are confined to the highest ranks. The surface of TIM is marked by a gash of high conservation, which covers half the interface and spreads over and around the ligand-buried site (Figure 3.1(TIM,c)). The remaining half of the interface is only moderately conserved and touches a nearby cluster of poor conservation. The progression described above is consistent with these observations and indicates that whereas patches that overlap the interface may score moderately, thanks to the intersection of the interface and the conserved gash, patches that cover the ligand-buried site and avoid the poorly conserved clusters score higher.

3.3.2 Significance of interface conservation

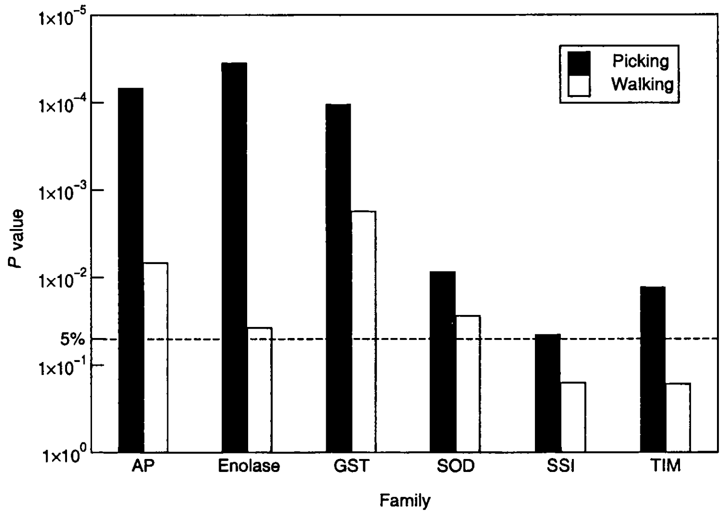
For a given protomer, the significance of interface conservation was assessed as follows. A set of residues equal in number to that of the interface was drawn at random from the surface 10 million times. The fraction of times this random set was at least as well conserved as the interface set was taken as the *P* value for the interface. If and only if the *P* value was less than the predefined cutoff 0.05, ie, such that the probability of interface conservation being random was <5%, the interface was considered significantly conserved. Two distinct random processes, picking and walking, were used to draw residues from the surface. *P* values were generated using both processes for both definitions of the interface in each family. Moreover, for each combination, *P* values were estimated in both the absence and presence of ligand-buried residues (see Methods for details). *P* values, being a relative measure, transcend absolute residues conservation. The use of *P* values therefore allows the meaningful comparison of conservation across families whose alignments may differ in the extent of their sequence diversity. Results of the *P* value estimations are presented in Table 3.2 and Figures 3.3 and 3.4; distributions of conservation for random selection are shown in Figure 3.5.

The picking simulations showed that all the interfaces studied, regardless of which interface definition was used or whether or not ligand-buried residues were excluded, were significantly conserved. Enolase consistently gave the lowest, ie, most significant, *P* values whereas the family with the highest *P* values depended on interface definition, SOD having the highest among total interfaces and SSI scraping just below the 5% cutoff among central zones (Table 3.2 and Figure 3.3).

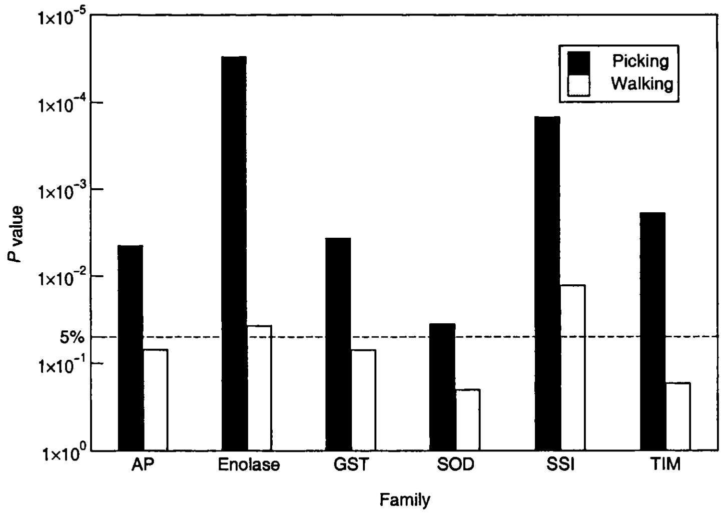
P values determined by walking were consistently higher than those determined by picking (Figure 3.4(a)), and in some cases were outside the top 5% of conserved walks. However, in every family except TIM the interface was significantly conserved by at least one of its two definitions (Table 3.2 and Figure 3.3).

The exclusion of ligand-buried residues (masking) affected *P* values by a small and usually negligible amount (Table 3.2). Its most conspicuous effect was seen in the walking *P* values for AP, and in particular those corresponding to the total interface, where masking promoted conservation of the interface from just outside the top 7% of walks to barely within the top 5%. Masking made only a small difference because, in most cases, the residues of a protomer's ligand-buried site numbered far fewer than those of its interface. The effect of their presence or absence in a pick or walk was therefore small.

Although absolute *P* values varied between picking and walking in a protomer, the rank order between one definition of the interface and the other did not (Figure 3.3). The central zone was unequivocally more

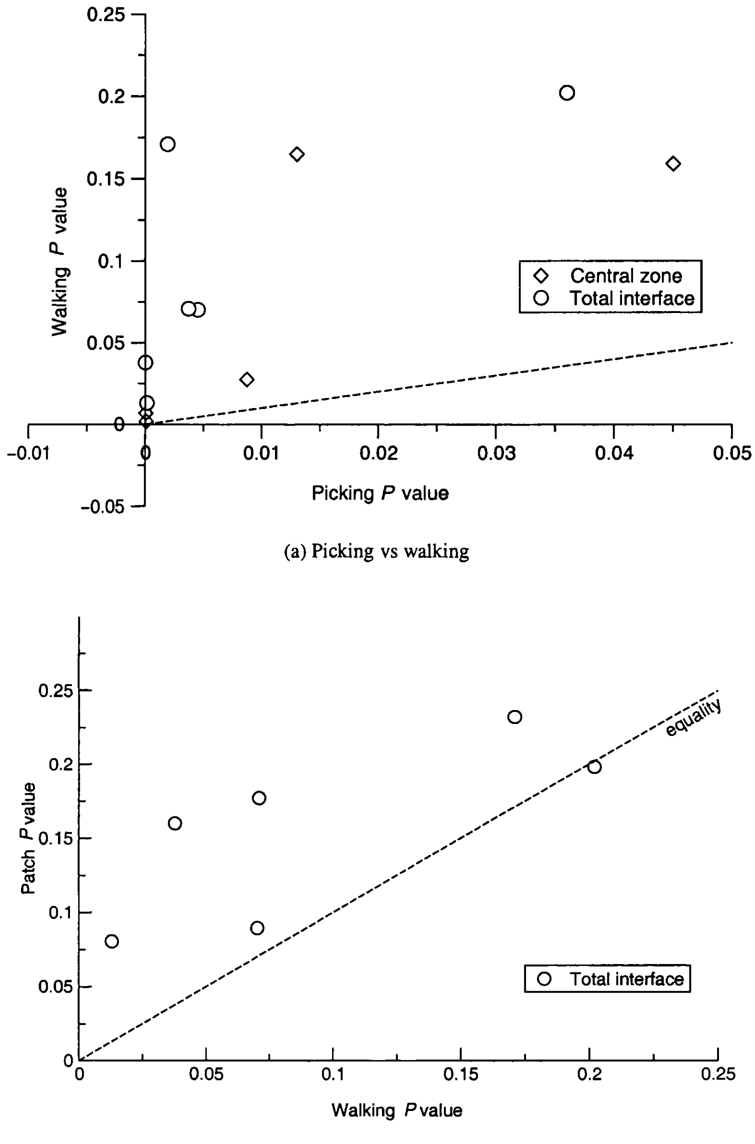


(a) Central zone



(b) Total interface

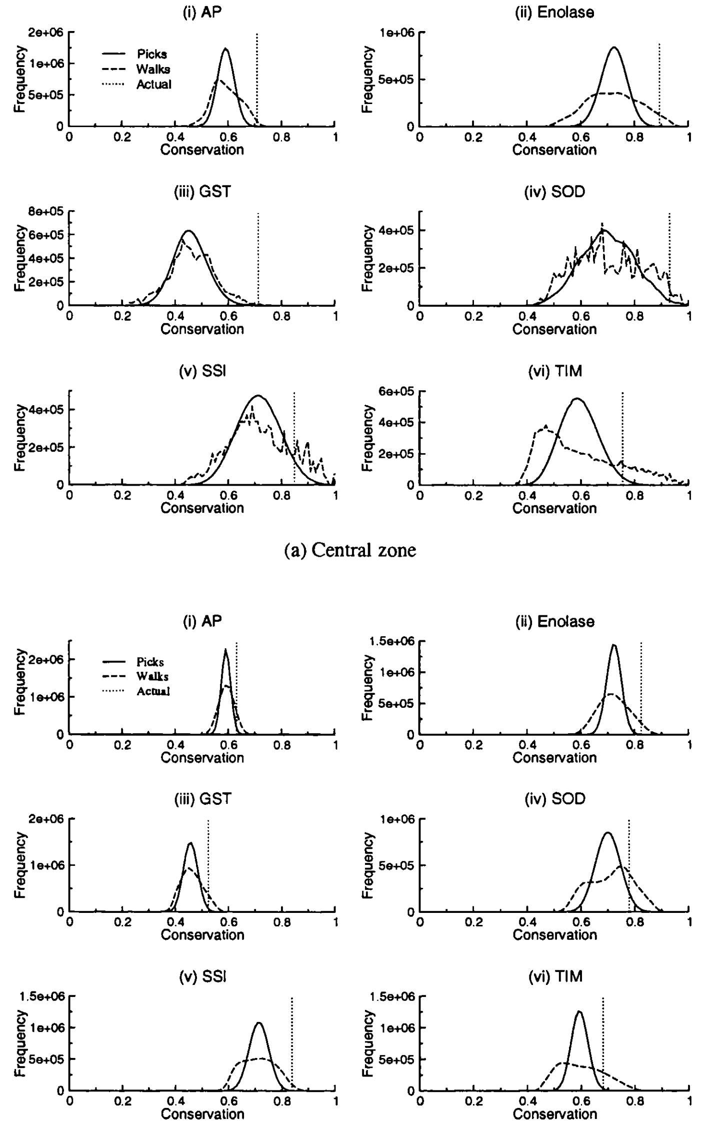
Figure 3.3: *P* values for residue conservation of the interface obtained by picking and walking for six families of homodimers.
Interfaces with *P* values smaller than 5%, ie, above the dashed line, are considered significantly conserved. *P* values are shown for two interface definitions: (a) central zone and (b) total interface. The graphs show *P* values from picking are significant for both definitions of the interface in all families, and that *P* values from walking are higher and significant less often. Note that *P* values are shown for the unmasked simulations only, since masking made negligible difference to these results.



(a) Picking vs walking

(b) Patch analysis vs walking

Figure 3.4: Comparison of *P* values obtained by different methods.
In graph (a) *P* values calculated by picking (x-axis) are plotted against *P* values calculated by walking (y-axis) for the same group of interface residues. The graph shows walking *P* values are consistently higher, ie, denote less significant conservation, than those calculated by picking and that there is little correlation between *P* values from the two methods. In graph (b), *P* values computed by walking (x-axis) are plotted against those generated from the patch analysis (y-axis) for the same interface. The graph shows walking *P* values typically give lower, ie, more significant, *P* values than from patch analysis and that *P* values from the two methods correlate reasonably. Note that *P* values are shown for the unmasked simulations only, since masking made negligible difference to these results.



(a) Central zone

(b) Total interface

Figure 3.5: Distributions of conservation for the picking and walking selection procedures. Each graph shows the distribution of conservation for 10 million picks (smooth line) and 10 million walks (dashed line) for one definition of the interface in one homodimer family. Graphs in (a) show distributions used to find *P* values for the central zone; graphs in (b) show distributions used to find *P* values for total interface. In each graph, conservation, defined as the mean conservation score of residues in a selection, is presented in bins of width 0.01 along the x-axis. The number of selections that fall into a given bin is presented on the y-axis. A dotted line intersecting with the x-axis indicates the mean conservation of residues in the true interface. The *P* value determined by a given simulation is the fractional area of its curve that falls to the right of the dashed line. The graphs show picking tends to give regular, normal curves whereas walking gives irregular and often highly skewed curves. The implications of these findings is discussed in the Results section. Note that distributions are shown for the unmasked simulations only, since masking made negligible difference to these results.

conserved than the total interface in AP, GST and SOD families. In Enolase and TIM, *P* values were similar between the two interface definitions. Only in SSI was the total interface clearly more conserved than the central zone.

3.3.2.1 Conservation in picks

Figure 3.5 shows distributions of the conservation scores achieved by picks and walks. As with the patch distributions, the mean of a picking distribution estimates the mean conservation of surface residues. The higher moments relate only to the shape of the underlying distribution of conservation for individual surface residues and the number of residues chosen. They say nothing about how conservation is dispersed over the surface since picking is blind to where residues are located in space.

Picks the size of the total interface fell into distributions that were narrower and more symmetric than picks the size of the central zone because they represent larger samples of the surface population and so show greater convergence toward the mean. Such properties are manifest in Figures 3.5(a,iv) and 3.5(b,iv): SOD's central zone, which comprises a mere five residues, has an wide and irregular picking distribution (Figure 3.5(a,iv)) whereas its total interface, which comprises 20 residues, has a distribution that is symmetric and regular (Figure 3.5(b,iv)). Owing to the greater spread of the distribution in Figure 3.5(a,iv), the absolute level of interface conservation required to reach significance is far higher for the central zone than for the total interface. Despite this, the central zone, which has absolute interface conservation at 0.93, achieves a more significant *P* value than the total interface (Table 3.2 & Figure 3.3), suggesting that evolutionary pressure to conserve residue type concentrates at the centre of the interface. In SSI, the converse is true: SSI's total interface achieves not only greater significance than its central zone but also a higher absolute conservation score (Table 3.2). Tamura et al (Tamura et al, 1995) have demonstrated the importance of Val13, a central zone residue, to dimer formation and overall stability in SSI. The results presented here suggest residues in the outer zone also play crucial roles in this regard.

The picking results for the total interface complement the results of the patch analysis, both giving an indication of the relative conservation of the interface. Patch analysis did not provide *P* values as such, but the probabilities of a patch chosen at random being more conserved than the interface were 0.09 (AP), 0.16 (Enolase), 0.18 (GST), 0.20 (SOD), 0.08 (SSI), and 0.23 (TIM). None of these patch *P* values are less than 0.05 and they correlate poorly (with a Pearson's correlation coefficient of 0.34) with the *P* values generated by picking. Some differences are particularly striking. For example, TIM achieved significant *P* values by picking but relatively high, ie, random, *P* values according to patch analysis. This is because picking is geometry-free and so escaped the effects of clustering that prevented TIM's interface from achieving a high patch score.

3.3.2.2 Conservation in walks

The walking distribution for a particular protomer and its interface was always more spread out than the corresponding picking distribution, indicating that the dispersion of surface conservation in all families was more clustered than random to varying degrees (Figure 3.5). This increase in distribution width marginalized the absolute conservation score of all the interfaces studied, pushing, in each case, a greater proportion of selections beyond the interface score. Thus the *P* values for walking were consistently higher that those for picking.

Figure 3.5 shows the walking distributions for each family. These are similar in shape to the distributions of patch scores shown in Figure 3.2 and corroborate inferences based on the patch data made above. The walking data give particularly strong support to these inferences because walking, unlike patch analysis, samples the space of every possible set of contiguous surface residues, which includes the interface set. The walking *P* values also correlate reasonably (with a Pearson's correlation coefficient of 0.77) with *P* values from the patch analysis, although they are significant more often (Figure 3.4(b)).

For TIM, the similarities between walking and patch data are conspicuous (Figures 3.2(f) and 3.5(b,vi)). TIM, the only protomer that failed to achieve a significant walking score for either definition of its interface, had walking distributions more positively skewed than those for any other family (Figures 3.5(a,vi), (b,vi)). Most skewed of all was its distribution for the central zone (Figure 3.5(a,vi)), where so many walks contained a majority of highly conserved residues that the resulting curve resembles an extreme value distribution. This resemblance, far from being coincidental, is a direct result of the surface composition of TIM, in which the gash of conservation (described above) relegates most walks that are outside it to the lower ranks of the distribution and promotes walks that overlap it to higher and higher ranks in diminishing numbers. TIM's walking results are thus consonant with its patch results and it is no surprise that its *P* values from both methods are the highest, ie, most random, among all families.

3.4 Discussion

The results show that the interfaces of all proteins studied here are more conserved than expected for a random distribution and that in most cases this conservation is statistically significant at the 5% level. In some cases, the selective pressure to remain invariant concentrates in the central zone; in others, conservation is about evenly matched across the complete interface; in one (SSI), selection against change may be strongest at the periphery of the interface.

Of the two methods used to select groups of surface residues, picking is the simplest and its results are the most straightforward to interpret. However, walking reveals more about the difficulties that would be inherent in predicting the location of interface using conservation alone. The strict definition of the ligand-buried site meant that many highly conserved residues that play a role in binding were ignored. If ligand-buried residues were defined as residues that lose merely some accessibility rather than those that become totally inaccessible on addition of ligand groups, the masked results would be different, probably giving lower *P* values in all cases except SSI, where the *P* values would be higher. However, it was felt that to exclude so many residues from the analysis would be misleading.

The results suggest the analysis methods described here could be usefully applied to the problem of differentiating crystalline contacts from biologically relevant interfaces (Ponstingl et al, 2000). Proteins crystallize as multimers that may contain both biological contacts, which are subject to evolutionary constraints, and non-biological contacts, which are not. If family information is available, picking or walking *P* values could be used to detect interactions in a crystal structure that are biologically relevant.

In this analysis we test whether conservation of an interaction is reflected in conservation of amino acid type at the site of that interaction. To ensure reliability, unusually stringent criteria were observed when compiling the dataset. For instance, it was compulsory that all sequences used in assessing conservation for a protomer share that protomer's multimeric states as explicitly recorded in their annotation. Further, in the interests of consistency, only homodimers were included: their annotation and nature of binding tend to be well documented, thus less likely to introduce confounding factors, than for other types of complexes. Such patterns may not be so distinct or be so readily detected for heterodimers or transient complexes. However, if the interfaces are functionally important, we expect them to be conserved. The challenge now is to use this information to help develop a method that can predict the location of an interface given only the structure of the protomer and its sequence alignment.

Chapter 4

Using conservation to identify biologically relevant crystal contacts

4.1 Introduction

Most crystal contacts are artifacts of crystallization that would not occur in solution or in the physiological state. But some of the observed contacts may be biologically relevant. Determining which contacts are biological and which are not is often difficult, particularly when, as frequently seems to be the case for entries in the Protein Data Bank (PDB) (Berman et al, 2000), the oligomeric state of the protein is uncertain or unknown (Henrick & Thornton, 1998, http://pqs.ebi.ac.uk/).

4.1.1 Biological contacts

Biological contacts, which here refer to any site of *in vivo* recognition between macromolecules, have received more attention than nonbiological contacts or comparisons of the two. Biological interfaces have been characterized in terms of their geometric features, such as planarity, shape-complementarity and circularity, in terms of their chemistry, such as hydrophobicity, preference for certain amino acids, and in terms of residue conservation (Chothia & Janin, 1975) (Janin et al, 1988) (Jones & Thornton, 1995) (Lijnzaad & Argos, 1997) (Lo Conte et al, 1999) (Valdar & Thornton, 2001). Although a number of studies have sought to predict the location of biological interfaces based on some of these parameters (Young et al, 1994) (Jones & Thornton, 1997b) or to dock partners (Sternberg et al, 1998, and refs therein), few have attempted to discriminate between biological and nonbiological contacts (Ponstingl et al, 2000), a problem faced by anyone who interprets X-ray data.

4.1.2 Nonbiological contacts

Most proteins solved by X-ray analysis and deposited in the PDB have three or more crystal contacts, and some have over 20. The sum of these contacts typically buries around 30% of the protein surface to ensure crystal stability (Carugo & Argos, 1997).

4.1.3 Comparing biological and nonbiological contacts

A number of features distinguish biological from nonbiological contacts. Biologically relevant interactions tend to be more specific than nonbiological ones, although this can be hard to detect in the crystal (Janin, 1997). The promiscuity of nonbiological contacts in pancreatic ribonuclease has been demonstrated by Crosio et al (Crosio et al, 1992). They showed that almost any residue on the surface of the protomer can be part of a crystal contact and that the same residue involved in two alternative contacts may interact with a different set of partners. Biological contacts tend to be larger than nonbiological ones and usually constitute the biggest contact in the crystal (Janin & Rodier, 1995) (Carugo & Argos, 1997) (Janin, 1997) (Dasgupta et al, 1997). The amino acid composition of nonbiological contacts is much like that of the surface as a whole (Carugo & Argos, 1997), although observed distributions vary slightly with ionic strength of solvent (Iyer et al, 2000). Biological contacts are split on the issue of composition. Transient contacts, such as those formed in signal transduction are composed similarly to the rest of the surface, whereas oligomeric contacts have a composition intermediate between the surface and the protein core (Jones & Thornton, 1997a) (Lo Conte et al, 1999). Some groups have mutated residues on the surface in order to engineer nonbiological contacts and so improve crystal stability (McElroy et al, 1992, for example).

4.1.4 Automatic discrimination of nonbiological from biological contacts

Automatic discrimination of biological from nonbiological contacts is desirable, and is attempted in the Protein Quaternary Structure database (PQS) (Henrick & Thornton, 1998). Because the contact size is such a powerful discriminant, the PQS uses accessible surface area (ASA) of the buried contact area to distinguish biological from nonbiological contacts, along with a number of other physical measures, which are not rigorously optimized. The method developed for PQS, when assessed against solution data for a nonredundant subset of proteins, distinguished correctly between true and false homodimers 78% of the time (Hannes Ponstingl private correspondence).

Ponstingl et al rigorously tested the utility of ASA and statistical "pair potentials" as discriminants (Ponstingl et al, 2000). Pair potentials are putative energies derived from a statistical analysis of observed frequencies of atom-pairs at a given separation. These have been used before for predicting the location of putative biological contacts (Robert & Janin, 1998) and for discriminating between computer-docked protein complexes (Moont et al, 1999). Ponstingl et al analysed a dataset of 172 proteins, with 76 homodimers and 96 monomers. Straight ASA produced a correct classification 84.6% of the time. Their pair potential correctly classified proteins in their dataset 87.5% of the time. A modified ASA score that considered the difference in size between the two largest contacts gave an accuracy of 88.9%.

Conservation has been used successfully to explore patterns of energy and define functional residues at protein binding sites (Lichtarge et al, 1996) (Lichtarge et al, 1997) (Lockless & Ranganathan, 1999) (Armon et al, 2001). Recently we reported that, within a small and extensively researched dataset, oligomeric interfaces exhibit significant residue conservation compared with comparable-sized regions of the protein surface (Valdar & Thornton, 2001). There is a clear rationale for why biological interfaces should be conserved: the amount by which they vary is circumscribed by the importance and specificity of their physiological role, and the degree of variability required to disrupt them. Conversely, we would expect no such selective evolutionary pressure on nonbiological contacts, which are the result of human experiments and not the product of evolution (Durbin & Feher, 1996). The above suggests conservation may be useful in discriminating between biological contacts, which we assume will be conserved, and nonbiological ones, which we assume will not. Moreover, since the measures of conservation and size are orthogonal, it is possible that combining them will provide a truly powerful discriminator.

We assess the utility of size and conservation in addressing the following two questions:

1. Is a given crystal contact biological?
2. Given all contacts in the crystal of a homodimer, which is the biological one?

These questions are different from those posed in earlier studies that have attempted to distinguish between homodimers and monomers. They more directly test the utility of conservation in identifying biological relevance of a contact. We develop algorithms that use one or both measures to answer each of the questions above. We compare efficacy of these algorithms, as well as the relative contribution of size and conservation to their predictive power.

4.2 Methods

4.2.1 Dataset

The dataset of Ponstingl et al (Ponstingl et al, 2000) was used to provide a starting point for further filtering. This comprised 172 non-homologous protein crystal structures of which 76 were homodimers and 96 were monomers. Atom coordinates were taken from the PDB. A program written by Hannes Ponstingl was used to generate hypothetical contacts for each structure. It works by applying crystallographic symmetry operations to a given protomer chain to recreate all contacts in the crystal. The term "protomer" is used here to denote a component chain of a multimeric complex.

In order to calculate residue conservation meaningfully, each protein must have a sufficiently large and diverse set of homologues. Because insufficient sequence information was available for some of the proteins in the initial dataset, the final dataset was smaller, comprising 118 proteins of which 53 were homodimers (1a3c, 1ad3, 1afw, 1ajs, 1alk, 1alo, 1amk, 1aom, 1aor, 1aq6, 1auo, 1bif, 1bsr, 1cg2, 1chm, 1cmb, 1cp2, 1csh, 1ctt, 1daa, 1fro, 1hjr, 1imb, 1isa, 1iso, 1kpf, 1lyn, 1mjl, 1mka, 1moq, 1nsy, 1oac, 1otp, 1pgt, 1pre, 1rfb, 1ses, 1slt, 1sox, 1tox, 1trk, 1tys, 1uby, 1wgj, 1xso, 2ilk, 2tct, 2tgi, 3grs, 3pgh, 3ssi, 4kbp, 5csm) and 65 were monomers (16pk, 1a0k, 1a6q, 1aay, 1af7, 1afk, 1ah7, 1ako, 1akz, 1am6, 1amj, 1aoh, 1aua, 1aun, 1avp, 1ayl, 1bc2, 1be0, 1bg0, 1bgc, 1bkz, 1bn8, 1bpl, 1bry, 1bwz, 1c3d, 1cki, 1dff, 1djx, 1dmr, 1esf, 1eso, 1fdr, 1feh, 1fsu, 1gci, 1inp, 1ips, 1kfs, 1mdt, 1mhl, 1mpg, 1pda, 1pjr, 1pmi, 1ppo, 1rgp, 1rhs, 1ton, 1uch, 1uro, 1xgs, 1yge, 1zin, 232l, 2atj, 2bls, 2fgf, 2ihl, 2mbr, 2pth, 2rn2, 3cms, 3sil, 8paz).

4.2.2 Definition of a contact

We consider only surface residues in our dataset. A residue is considered to be on the surface if its relative accessible surface area (RSA; (Lee & Richards, 1971)) in the isolated protomer is greater than 5% of the maximum for an extended tripeptide in which that amino acid is flanked by alanines. If the residue's RSA is less than 5%, it is considered part of the structural core of the protein. Solvent accessibility was determined using NACCESS (Hubbard & Thornton, 1993), an implementation of the Lee & Richards algorithm (Lee & Richards, 1971), with a probe sphere of radius 1.4Å. The surface cutoff used follows that devised by Miller et al (Miller et al, 1987).

A given protomer in the dataset is surrounded by a number of partners. Each partner touches the protomer surface, defining a different crystal contact. A contact is described by the set of residues on the surface of the protomer that each lose at least 1 Å^2 of ASA when complexed with the relevant partner. Crystal contacts that fail to bury any residues by this amount are excluded from the dataset. A given surface residue may therefore be classified as one of the following: part of a biological contact, part of a nonbiological contact or belonging to the rest of the surface.

4.2.3 Assessing conservation

4.2.3.1 Identification and alignment of homologues

Homologues were identified for each protomer from the Non-Redundant DataBase (NRDB, a database of protein sequences maintained by the NCBI) using the iterative profile search program PSI-BLAST (Altschul et al, 1997). PSI-BLAST was allowed a maximum of 20 iterations to reach convergence. The E-value threshold for inclusion of new homologues at each iteration was set at 10^{-40}. This cutoff was strict enough to guard against profile drift but sensitive enough to allow detection of remote homologues, with the sequence identity of a match to the query falling as low as 5%. The profile alignment used in a protomer's final PSI-BLAST run was taken as the multiple alignment for that protomer. Multiple alignments comprising fewer than four sequences, including that of the protomer query sequence, were regarded as containing insufficient evolutionary information and were excluded.

4.2.3.2 Scoring residue conservation from an alignment

A score of evolutionary conservation, ranging continuously between 0 for unconserved and 1 for strictly conserved, was assigned to each residue in the protomer from its multiple alignment using the *Cons* sum of pairs score described by Valdar & Thornton (Valdar & Thornton, 2001). *Cons* uses amino acid similarities inferred from PET (Jones et al, 1992), a Dayhoff-like mutation data matrix (Dayhoff et al, 1978), to assess the diversity of amino acids at an aligned position. In this score, contributions from individual sequences are weighted inversely with their redundancy in the alignment.

4.2.3.3 Probabilistic scoring of contact conservation

Contact conservation was scored probabilistically after the "picking" measure described in chapter 3 (section 3.2.7.1). In this scheme, the conservation of a contact of size m is described by $P_{Cons}(Cons, m)$, the probability that a group of m residues drawn at random without replacement from the surface of the protomer has an average *Cons* score greater than or equal to that of the m residues in the contact. This probability is computed by simulation: m residues are chosen at random from the surface and their mean *Cons* recorded. This is repeated 1 million times to give a probability estimate with an expected error of at most 10^{-3} at least 95% percent of the time. Similar P values could have been computed using simpler statistical tests that do not require simulation, eg, the Z-test. However, the small size of some contacts was felt to undermine the assumptions made by such tests, making simulation the more robust alternative.

Low values of P_{Cons} denote highly conserved patches, reflecting that such high average residue conservation would be unlikely from a chance draw. High values of P_{Cons} denote poor conservation, likely to be bettered in a chance draw.

4.2.3.4 Filtering the dataset for uninformative cases

If sequences in a family are too similar, conservation becomes uninformative. For instance, consider a protomer with 100 surface residues; 99 have a *Cons* of 1 and one has a *Cons* of 0.5. A ten-residue contact on that surface could achieve one of only two possible P_{Cons} scores: 1 if it contained the less conserved residue and $\simeq 0.57$ otherwise. This kind of granularity is undesirable; after all, it is grossly misleading for a contact of maximal conservation to receive a P value above 0.5. To filter out cases in which the diversity of surface scores is not sufficient to support a meaningful P value, we apply to each protomer surface a function *Doss* (*D*iversity *o*f *s*urface *s*cores) and reject cases that fall below a cutoff for this metric.

Although the actual diversity of P_{Cons} scores is complex to work out, varying both with protomer and contact size, a simple count of the number of distinct permutations of residue scores indicates this property well enough. A natural measure is thus the multinomial coefficient, conveniently expressed in Shannon's entropy (Shannon, 1948). For a given protomer, let *All* be the set of *Cons* scores belonging to all surface residues and *Unique* be the non-redundant set of these scores. Then the diversity of surface scores, *Doss*, is given by

$$Doss = \left[- \sum_{i \in Unique} \frac{n_i}{n(All)} \log \frac{n_i}{n(All)} \right] \times \frac{100\%}{\log n(All)},$$

where n_i is the number of instances of score i, and $n(All)$ is the number of surface residues. *Doss* ranges between maximal diversity at 100% and uniformity at 0%.

Applying *Doss* to the original dataset resulted in a distribution of three parts: a minority of protomers occupied the ranges 0-25% and 35-65%, whereas the majority sat in the range 70-100%. Because alignments in the lowest range were perceptibly redundant, a cutoff of 30% *Doss* was chosen and all protomers with alignments falling below this threshold were excluded.

4.2.3.5 Testing whether conservation of biological contacts is significant

To test whether biological interfaces are usually more conserved than nonbiological contacts, we first count how many times the most conserved contact around a protomer is biological. Second, we compute the probability ($P_{mostcon}$) of observing such a result with a null model in which all contacts, regardless of type, are equally likely to be the most conserved. Last, the value of $P_{mostcon}$ is used to assess how well the null model accommodates the observed results and to infer whether the frequency with which biological contacts are most conserved is statistically significant.

A null model is proposed that assumes biological contacts are no more or less conserved than nonbiological ones. According to this model, the most conserved contact of the ith protomer is a random draw on the n_i contacts that surround it. Let "success" describe an event in which the most conserved contact is biological. The probability of a success in protomer i is then given by the Bernoulli distribution f_i:

$$f_i(X) = \begin{cases} 1 - \frac{1}{n_i} & \text{if } X = 0 \text{ (failure)} \\ \frac{1}{n_i} & \text{if } X = 1 \text{ (success)} \end{cases}.$$

Across $N = 53$ homodimers, the total number of successes depends on all N distributions and has a probability mass function $h = f_1 * f_2 * \ldots * f_N$, where $f_i * f_j$ is the convolution of distributions f_i and f_j. If the most conserved contact is observed to be biological m times, then probability of the null model achieving at least m successes is given by

$$P_{mostcon}(m) = \sum_{i=m}^{N} h(i).$$

	Absolute measures	Relative measures
Size	number of residues in the contact	ranked size among a protomer'scontacts (measured as thefractional rank[a], such that thebiggest contact has a rank of1.0 and smaller contacts haveranks of <1)
	fraction of surface residuesburied by the contact	size difference from thelargest contact in the protomer(eg, if the biggest contactburied 20 residues, a contactthat covered only 15 wouldhave a size difference of $20 - 15 = 5$)
Conservation	P_{Cons} for the contact(ranges from 0, meaning highlyconserved, to 1, meaningpoorly conserved)	ranked P_{Cons} among aprotomer's contacts (measuredas the fractional rank[a], such thatthe most conserved contact hasa rank of 1.0, and less-conserved contacts have ranksof <1)

Table 4.1: Measures available to predictors

[a] Fractional rank is the rank of an item divided by the maximum rank for that set of items. Fractional rank takes values in the range (0,1].

4.2.4 Discriminating biological from nonbiological contacts

We examine the discrimination problem from two viewpoints, addressing the following questions:

1. Absolute assessment: is a given contact biological?
2. Relative assessment: given the set of contacts associated with a protomer that is known to be homodimeric, which contact among this set is biological?

We devise a number of predictors to answer these questions using the size and conservation data available. The absolute assessment pools crystal contacts from a set of protomers. A predictor attempts to classify each contact as biological or nonbiological. This assessment is performed on two sets of contacts: the homodimer set, which comprises both types of contact, and the monomer set, which contains only nonbiological contacts. The relative assessment is performed on the homodimers. Predictors consider each protomer in turn, deciding which of its contacts is biological. A prediction is correct if the true biological contact only is classified as biological. If other contacts are classified as biological in addition to or instead of this contact, the prediction is deemed incorrect.

Table 4.1 shows the measures available to the predictors. Predictors in the absolute assessment may use only absolute measures whereas predictors in the relative assessment may use absolute or relative measures.

4.2.4.1 Neural network predictors

Neural networks can provide an elegant and convenient framework for classifying new data based on patterns extracted from old data. Herein, we use two types of feed-forward neural network: the single layer perceptron (SLP) and the two-layer multilayer perceptron (MLP) (Wu & McLarty, 2000, and refs therein). The inputs are a selection of the measures listed in Table 4.1 and the output corresponds to the predicted class: biological or nonbiological. Multilayer perceptrons may contain different numbers of hidden units, so for notational convenience if an MLP has x hidden units, it is referred as MLPx (eg, MLP2 denotes an MLP with two hidden units).

The neural networks employed here are trained using the scaled conjugate gradient algorithm (Møler, 1993), which is usually faster than the more traditional backpropagation. The number of training iterations was set constant at 100.

4.2.5 Assessing discriminator performance

4.2.5.1 Cross-validation

We assess the performance of the neural network predictors using a "leave-one-out" form of cross-validation (jackknifing). In this scheme, for a dataset of n protomers, a predictor is trained on data from n-1 protomers and tested on data for the single remaining protomer. This is then repeated for each protomer in turn.

Neural network predictors are jackknifed for the absolute and relative assessments on the homodimers. In the absolute assessment on the monomers, which contains no positive examples, networks train on the whole homodimer set.

4.2.5.2 Performance measures

Performance of predictors in the absolute assessment is measured in three ways: by accuracy, error rate and a comparison with random. All three scores can be derived from the following quantities:

p = number of correctly classified biological contacts

n = number of correctly classified nonbiological contacts

o = number of nonbiological contacts classified as biological (overpredictions)

u = number of biological contacts classified as nonbiological (underpredictions)

t $= p+n+o+u$

The most straightforward score, "accuracy" measures the percentage of correctly classified contacts:

$$\text{accuracy} = \frac{p+n}{t} \times 100\% .$$

The "error rate" measures the percentage of incorrectly classified contacts and is simply 100% minus the accuracy.

Accuracy and error rate can be misleading when the dataset contains many more instances of one class than another. For instance, consider a predictor that has no discriminatory power and just predicts everything to be nonbiological. Because the vast majority of contacts in the homodimer dataset are nonbiological, this predictor would automatically give high accuracy. To penalize such spurious achievements, we include a third score, the phi-coefficient (hereinafter referred to simply as "phi").

Phi (also referred to as "Matthew's correlation coefficient" in neural network literature) measures the correlation between observed and predicted results. It is a special case of Pearson's correlation coefficient, computed when the two variables being compared are dichotomous and take values of 0 or 1. Phi ranges from -1, representing inverse correlation and extremely poor predictive power, to +1, representing perfect

		Homodimers	Monomers
Number of protomers		53	65
Number of contacts	biological	53	0
	nonbiological	366	535
	total	419	535
Number of contacts per protomer	range	[3, 14]	[4,13]
	mean	7.9	8.2
Number of sequences in alignment	range	[4, 251]	[6, 251]
	mean	72.4	82.8
Doss score[a]	mean	82.5%	84.0%
	s.d.	17.5%	18.0%

Table 4.2: Summary statistics for the homodimer and monomer datasets

[a] *Doss* measures the diversity of conservation scores on the surface of the protein. Low percentages indicate the alignment may not be diverse enough for meaningful conservation scores to be extracted.

correlation and an ideal prediction. Phi is calculated as

$$\phi = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} .$$

In addition to possessing a convenient range, phi also has a real probabilistic basis, relating to the chi-squared function for 2 x 2 contingency table. Specifically, it can be shown that $\phi = \sqrt{\chi_1^2/t}$, which means the likelihood of random assignment producing the observed prediction rate, $p_{\chi_1^2}$ (the chi-squared probability), is one at $\phi = 0$ and decreases with increasing phi (Sheskin, 2000).

Accuracy and phi were used to measure performance of predictors applied to the homodimers. Performance on the monomers was assessed using the error rate. A prediction in the relative assessment is either right or wrong: there are no true negative examples. We measure predictor performance by accuracy, calculated as the percentage of correct predictions.

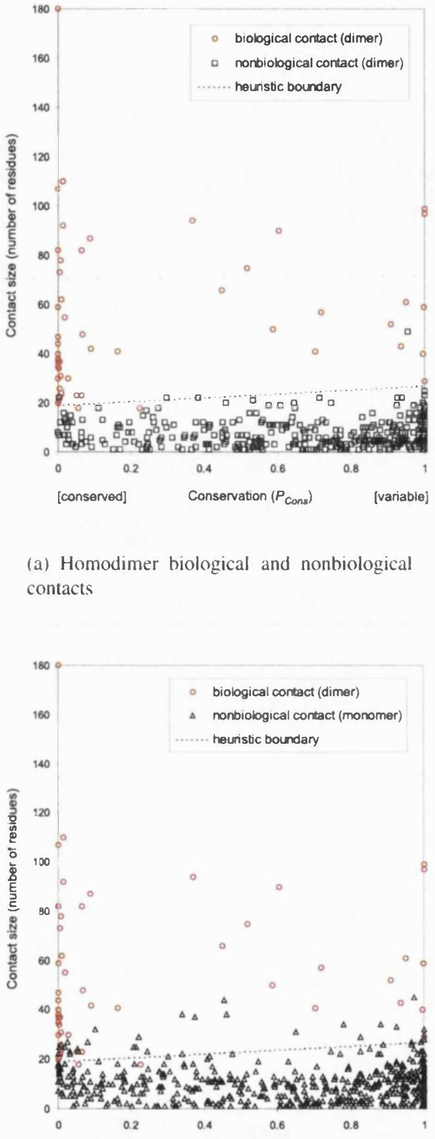
4.3 Results

We investigated size and conservation of crystal contacts in 53 families of homodimers and 65 families of monomers. A contact was defined as the set of residues on a protomer that lose their accessibility upon complexation with a partner. Contact conservation was measured probabilistically as P_{Cons}. On this scale, values close to zero indicate extremely high conservation (ie, improbable by chance) and values close to one indicate extreme low conservation (ie, high variability in evolution). Table 4.2 shows the number of biological and nonbiological contacts in each dataset, and information about the family alignments. Figure 4.1 plots the size of contacts from each set against their conservation.

4.3.1 Contact size

Biological contacts were typically bigger than nonbiological contacts. The average biological contact was 53.7 residues in size and accounted for 25.9% of a protomer's surface residues. In contrast, the average nonbiological contact was a mere 7.6 residues and covered only 4.2% of the surface.

Figure 4.2 shows distributions of contact size in the dataset. These distributions reveal substantial variation in size among both types of contact but particularly for biological ones. A significant number of biological contacts occupy an area in the lower ranks of the distribution that overlap with high numbers of



(a) Homodimer biological and nonbiological contacts

(b) Homodimer biological contacts and monomer nonbiological contacts

Figure 4.1: Size and conservation of crystal contacts in the homodimers and monomers. Size is plotted as the number of residues. Conservation is plotted on the P_{Cons} scale, where 0 is highly conserved and 1 is highly variable (unconserved). Graph (a) plots size against conservation for biological (red circles) and nonbiological (black squares) crystal contacts in the homodimers. Graph (b) plots these measures for nonbiological contacts in the monomers (black triangles) and, for comparison, plots biological contacts in the homodimers (red circles). In each graph, a dotted line represents the decision boundary devised for the heuristic predictor H_{abs} (see Results), which attempts to automatically separate biological from nonbiological classes of data based on size and conservation. These graphs show that the two classes naturally separate by size and, to a lesser extent, by conservation.

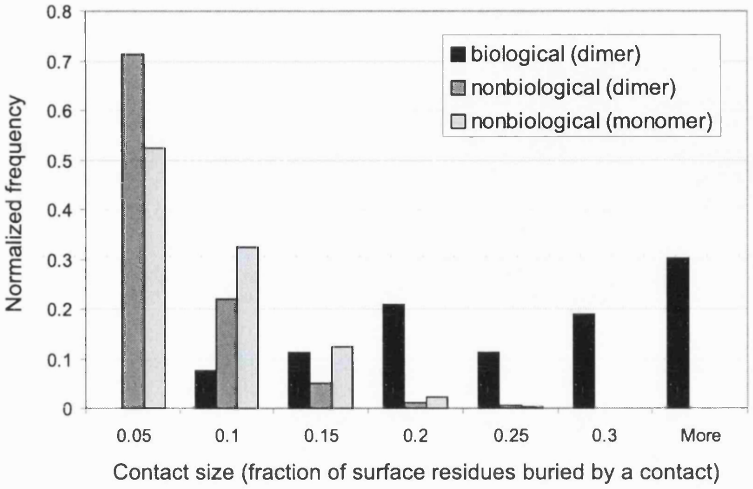


Figure 4.2: Size of contacts in the homodimer and monomer crystals.
Contact size, measured as the fraction of surface residues buried in the interface, is presented on the x-axis in bins of 0.05 from 0 to 0.3 and by a single aggregate bin, labelled "more", thereafter. The distributions show that biological contacts tend to be larger than nonbiological contacts, but that there is a significant region of overlap between these two classes.

nonbiological contacts.

The biological contact was the largest contact made by the protomer in all but one of the 53 homodimer crystals. The exception was 1uby, in which the 48-residue biological contact took second place to a 49-residue nonbiological one.

4.3.2 Contact conservation

Biological contacts were usually more conserved than nonbiological ones (Figure 4.3). On average, biological contacts had a P_{Cons} score of 0.26, whereas nonbiological contacts scored an average of 0.67 in the homodimers and 0.63 in the monomers. The biological contact was the most conserved contact surrounding the protomer in 36 of the 53 homodimer crystals. The calculated $P_{mostcon}$, which describes the probability of this happening in a null model where the most conserved contact is a random draw (see Methods), was 2.38×10^{-19}.

Despite these figures, biological contacts were not exclusively highly conserved and highly conserved contacts were not exclusively biological. Figure 4.3 plots the distribution of conservation for the two contact types in the homodimer and monomer sets. It shows that although biological contacts tend to be conserved, these contacts exhibit a full range of conservation, with the second most frequent group at the least conserved extreme. The distributions for nonbiological contacts in the homodimers and monomers are strikingly similar. In both, the mode coincides with extreme evolutionary variability whereas the remaining contacts span the range of P_{Cons} about evenly. Figure 4.4 uses Bayes theorem to combine the distributions for homodimers in Figure 4.3. It consolidates the above findings: the likelihood of a contact being biological diminishes as its residues become more variable in evolution. Figure 4.5 consolidates the relationship

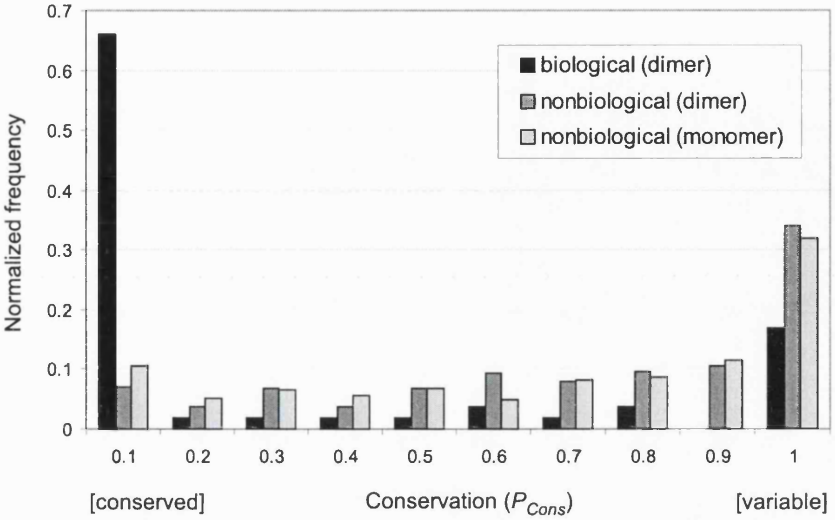


Figure 4.3: Conservation of biological and nonbiological contacts in the homodimer and monomer crystals. Conservation is measured on the P_{Cons} scale (0=strongly conserved, 1=poorly conserved). Conservation is presented in bins of width 0.1 on the x-axis. Frequency, measured as the fraction of data belonging to a given class that falls into that conservation bin, is presented on the y-axis. The histograms show biological contacts tend to be highly conserved whereas nonbiological contacts tend to be poorly conserved.

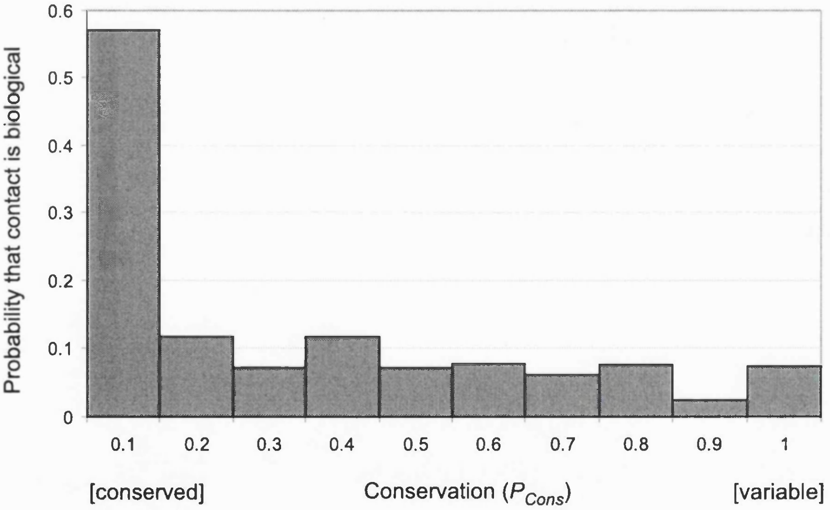


Figure 4.4: Probability of a contact being biological given its conservation in the homodimers. Conservation is measured on the P_{Cons} scale (0=strongly conserved, 1=poorly conserved). Conservation is presented in bins of width 0.1 on the x-axis. The height of each bar represents the probability that a contact selected at random is biological given that it has the conservation associated with its x-axis bin. These probabilities are computed according to Bayes theorem. To avoid zero probabilities, constant pseudocounts of 1 were added to raw frequencies (according to Laplace's rule) before Bayes theorem was applied. The histogram shows that the probability of a contact being biological sharply decreases with decreasing conservation.

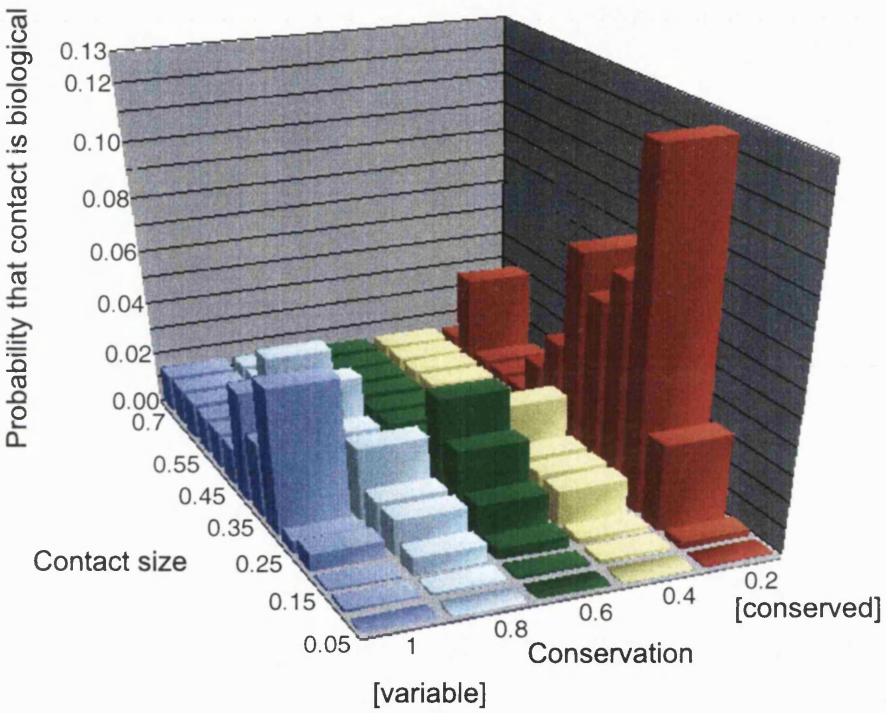


Figure 4.5: Probability of contact being biological given its size and conservation. Contact conservation, as P_{Cons} (0=strongly conserved, 1=poorly conserved), is presented in bins of width 0.2 on the x-axis. For clarity, each conservation range is depicted in a different colour. Contact size, measured as the fraction of surface residues buried by a contact, is presented on the z-axis in bins of 0.05 from 0 to 0.7. The y-axis (vertical) measures the probability of a contact, randomly selected from the dataset, being biological given its size and conservation. Probabilities are calculated according to Bayes theorem after first applying constant pseudocounts according to Laplace's rule. The graph shows that although larger contacts are more likely to be biological than smaller ones, high contact conservation makes this even more likely. The graph tails off around the higher values of size because there is little data for contacts in this range.

between size and conservation in Figure 4.1, showing that although highly conserved contacts are likely to be biological, poorly conserved contacts may also be biological provided they are large.

Conservation and size did not correlate significantly. P_{Cons} vs the number of residues in a contact gave a Spearman's rank-order correlation coefficient (Press et al, 1996) of -0.14. P_{Cons} vs the fraction of surface residues buried in a contact gave a similarly insignificant correlation of -0.17.

4.3.2.1 Poorly conserved biological contacts

Some biological contacts were extremely poorly conserved. At the level of the alignment, a biological interface may achieve a poor conservation score either because residues in the contact vary considerably or are subject to deletions. To investigate this we devise a score, Gappyness, that measures the extent to which the biological interface coincides with gaps in its multiple sequence alignment. If *Interface* is the set of positions in the alignment corresponding to residues in the biological contact of the protomer, $Gaps_i$ is the set of gaps aligned to the target sequence at position i, and $Aminos_i$ is the set of residues aligned at this position, then Gappyness is defined as

$$G = \sum_{i \in Interface} \frac{n(Gaps_i)}{n(Gaps_i \cup Aminos_i)} \times 100\%,$$

where $n(A)$ denotes the number of elements in set A. Gappyness for a protomer ranges from 0%, denoting no gaps aligned to the interface, to 100%, denoting that only gaps are aligned to the interface. The homodimers had an average Gappyness of 8.4% with a standard deviation of 10.3%. Gappyness is one of many possible causes of low interface conservation and so, unsurprisingly, Gappyness showed no significant correlation with P_{Cons} scores.

At the level of a homologous family, a biological interface may not be conserved because other members of its family are not homodimers, other members of the family are homodimers but dimerize in a different way, or because other members of the family are homodimers but variability at the interface confers multiple binding specificity in that family.

For the nine least conserved biological interfaces, Table 4.3 lists Gappyness and evidence for multiple multimeric states (MMS) within the family.

Poor interface conservation of 1alo ($G = 47.0\%$) and 1tox ($G = 38.6\%$) coincides with high Gappyness. 1alo is the crystal structure of aldehyde oxidoreductase extracted from *Desulfovibrio gigas* (Romão et al, 1995). 1alo, often referred to as MOP, is a member of the molybdenum hydroxylase family of enzymes (Hille, 1999). Its 149-sequence alignment contains many other members of this family, most of which, judging by the available annotation, are likely homodimers. Yet despite their common multimeric state, more than half of these homologues lack MOP's N-terminal tail, which for MOP constitutes a substantial portion of the interface. In the alignment there are at least two subfamilies: the aldehyde oxidoreductases (AO), which include a MOP-like N-terminus, and the xanthine dehydrogenases (XDH), which do not. Figure 4.6 shows that although XDH and AO are both dimers, the XDH protomer binds its partner in a quite different manner with a different part of its equivalent surface.

1tox is the crystal structure of diphtheria toxin extracted from *Candida albicans* (Bell & Eisenberg, 1996). 1tox comprises three domains, each with a separate function: a catalytic domain (C) at the N-terminus, a translocation domain in the middle (T) and a receptor binding domain (R) at the C-terminus. The 21 sequences in the 1tox alignment fall into three groups: those with all three domains, those that possess only domains C and T, and those with domain C only. The missing domains

Protomer	Name	P_{Cons}	Gappyness[a]	Multiple Multimeric States detected in Swissprot sequences[b]
1ad3	Aldehyde dehydrogenase (rat liver)	1	15.4%	18 / 29
1alo	Aldehyde oxidoreductase (*D. gigas*)	0.996	47.0%	0 / 10
1aor	Aldehyde ferredoxin oxidoreductase (*P. furiosis*)	1	7.9%	1 / 2
1bsr	Ribonuclease A (bovine seminal plasma)	0.936	4.3%	0 / 3
1nsy	NAD synthetase (*B. subtilis*)	0.997	5.4%	0 / 2
1slt	Galectin (bovine spleen)	1	15.1%	0 / 4
1tox	Diphtheria toxin (*C. albicans*)	1	38.6%	0 / 2
2tct	Tetracycline repressor (*E. coli*)	0.950	4.6%	0 / 1
5csm	Chorismate mutase (yeast)	0.909	10.6%	0 / 1

Table 4.3: Gappyness of interface for least conserved biological contacts ($P_{Cons} > 0.9$)

[a] Gappyness measures the extent to which gaps in the alignment coincide with the interface (see Results).

[b] Data is in the form n/N, where N is the number of Swissprot sequences in the alignment that have subunit annotation and n is the number of those sequences that are annotated as being something other than homodimer.

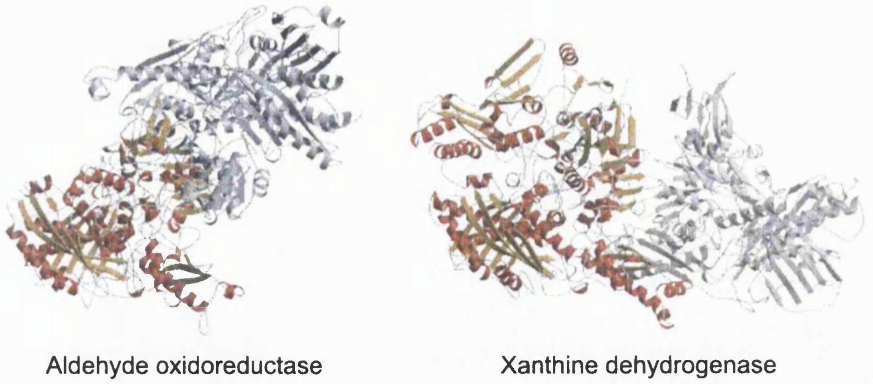


Figure 4.6: Comparison of homodimerization in aldehyde oxidoreductase (AO; left), and xanthine dehydrogenase (XDH, right).
These two proteins are homologous homodimers. The two protomers coloured red and gold are shown in equivalent orientations. The protomer partner of each is coloured in grey. The figure shows that despite homology and identical multimeric states, the mode of binding in these two homodimers is different. AO and XDH are represented by PDB (Berman et al, 2000) structures 1alo (Romão et al, 1995) and 1fiq (Enroth et al, 2000) respectively. Images were created using MOLSCRIPT (Kraulis, 1991) and Raster3D (Merritt & Bacon, 1997).

bestow high Gappyness, but the variability within aligned domains is negligible. The *Doss* score for 1tox reflects this. *Doss* measures the diversity of conservation scores in an alignment (see Methods). A low *Doss* score, such as that of 1tox (*Doss* = 34.9%), suggests the alignment may not be sufficiently diverse to support meaningful analysis of conservation. Examining the available annotation for sequences in the 1tox alignment reveals that many of the homologues lacking the R or T domain are fragments. This in turn suggests much of the observed variation in the alignment is indeed spurious.

Moderate Gappyness was seen in 1ad3 ($G = 15.4\%$) and 1slt ($G = 15.1\%$). 1ad3 is the crystal structure of aldehyde dehydrogenase extracted from rat liver (Liu et al, 1997). It comprises three domains: one NAD-binding, one catalytic and one bridging. The 1ad3 interface is large and elaborate, and involves all three domains. The bridging domain, which is believed to be important for stabilizing the 1ad3 dimer, is absent from the vast majority of the 251 sequences in this protomer's family. The nine interface residues that lie within this domain are, as a result, almost unmatched in the alignment and blight 1ad3 with its Gappyness and diminished overall conservation. The existence of MMS in the family further suggests the dimeric nature of 1ad3 is not important for many other members of its family.

1slt is the crystal structure of galectin, also known as S-type lectin, extracted from bovine spleen (Liao et al, 1994). The alignment of 1slt has gaps spread thinly throughout, rather than concentrated in a few conspicuous regions. It is likely that 1slt's poor conservation results from multiple multimeric states. Although the 1slt alignment contains few annotated sequences to support this (Table 4.3) preliminary runs of PSI-BLAST at lower inclusion E-values (eg, 0.0005) matched a large number of non-homodimeric sequences. Moreover, both dimers and tetramers occur at the level of 1slt's homologous superfamily in CATH (2.60.120.60). Lectins are a group of carbohydrate-binding proteins that exhibits a diverse range of structure and specificity, in which heterogeneity of quaternary structure is common (Vijayan & Chandra, 1999, and refs therein). MMS is also a likely cause of variability at the interface in 1aor (Table 4.3).

Neither Gappyness nor MMS were present in 1bsr. 1bsr is the crystal structure of ribonuclease A extracted from bovine seminal plasma. It is often referred to as BS RNase (Mazzarella et al, 1993). BS RNase is considered something of an outlier among ribonucleases, being the only surviving member of the seminal plasma RNases. Seminal plasma RNases are thought to have arisen from the same gene duplication event that spawned pancreatic and brain RNases in mammals (Sasso et al, 1999). These three families are paralogues: they are homologous but have diverged in function; and whereas both pancreatic and brain RNases have many active orthologues (homologues with equivalent function), BS RNase has none. The alignment of 1bsr thus contains many such paralogues, which are under evolutionary pressures different from those on BS RNase, and the only orthologues are engineered versions of 1bsr. Little information exists about the structure of brain RNase, but it is clear that the interfaces of pancreatic and BS RNase are different. In both cases, protomers intertwine termini to form metastable domain-swapped dimers (Bennett et al, 1995, and refs therein). However, whereas in BS RNase the dimer association is obliged by a disulfide bond, in pancreatic RNase this constraint is absent and dimer association takes second place to a more stable monomer form.

Candidate reasons for poor interface conservation of 1nsy, 2tct and 5csm were not found.

4.3.2.2 Highly conserved nonbiological contacts

Some nonbiological contacts were extremely highly conserved. Twenty-six (17%) nonbiological contacts achieved $P_{Cons} < 0.1$. All of these were relatively small, covering less than 10% of the surface of their parent protomer.

The most obvious explanation for why a small contact that is not a biological oligomeric is conserved

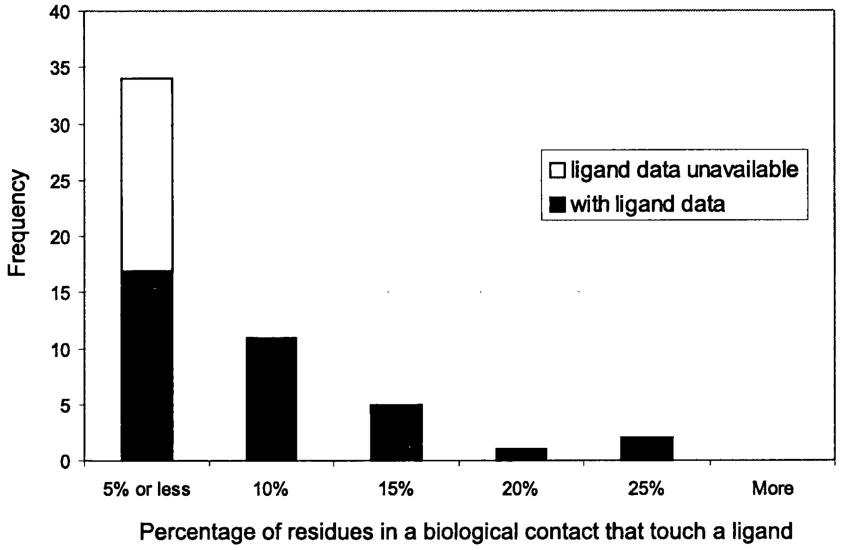


Figure 4.7: Overlap of biological contacts with ligand-touching residues.
Overlap on the x-axis is defined as the percentage of residues in a biological contact that are designated as ligand-touching (see Results). Ligand data was available for only 36 of the 53 homodimers. The graph shows that when overlap with a ligand-binding site occurs, it is usually moderate.

is that it overlaps a site of substrate or cofactor binding (Valdar & Thornton, 2001). For each protomer we identified "ligand-touching" residues (ie, surface residues that lose more than 1Å^2 ASA upon inclusion of the PDB ligand) and deduced which contacts contained them.

Ligands were present in 36 of the 53 homodimers, providing information for 36 biological and 231 nonbiological contacts. Figure 4.7 shows that for most of the 36 biological contacts, fewer than 10% of their residues are ligand-touching.

Ligand-touching residues significantly increased the conservation of some nonbiological contacts. However, they had an insignificant impact on the conservation of biological contacts. This can be shown from the data in Table 4.4. Among nonbiological contacts, the presence of ligand-touching residues was rare and high conservation rarer still. Given the frequencies of these two properties, the number of contacts observed with both (ie, 7) is greater than expected (expected = 19 x 35/231 = 2.9). The converse is true for biological contacts. Among these contacts both high conservation and the presence of ligand-touching residues are common. The number of biological contacts with both is 21, which is much as expected by chance (expected = 25 x 29/36 = 20.1). Fisher's Exact Test (Fisher, 1990) (see Table 4.4) confirms that the association between high conservation and the presence of ligand-touching residues is statistically significant at the 5% level for nonbiological contacts, but not significant at anywhere near this level for biological contacts. Therefore, sites of ligand-binding can lead to conservation misclassifying nonbiological contacts as biological.

	Biological (total 36)	Nonbiological (total 231)
highly conserved[a]	25	19
contains ligand-touching residues	29	35
highly conserved and contains ligand-touching residues	21	7
P_{Fisher}[b]	0.359	0.0131

Table 4.4: Contacts containing ligand-touching residues and contacts with high conservation in the homodimers.

[a] $P_{Cons} < 0.1$

[b] Probability from Fisher's Exact Test. Low values indicate a statistically significant association between high conservation and overlap with ligand-touching residues (see text for discussion). For a given type of contact, the probability of the association is calculated from the hypergeometric distribution as

$$P_{Fisher} = \sum_{k=n_{both}}^{n_{cons}} \left(\binom{n_{cons}}{k} \binom{n_{total} - n_{cons}}{n_{ligand} - k} \Big/ \binom{n_{total}}{n_{ligand}} \right),$$

where n_{total} is the total number of contacts of that type, n_{cons} is the number of these that are highly conserved, n_{ligand} is the number of these containing ligand-touching residues and n_{both} is the number that are both highly conserved and contain ligand-touching residues.

4.3.3 Discriminatory power of size and conservation

4.3.3.1 Heuristic predictors

We devised the following simple heuristic predictors from a visual inspection of the raw data. The heuristic predictor for absolute assessment, H_{abs}, traces a straight line on a graph of contact size against P_{Cons} (such as in figure 4.1), separating biological from nonbiological contacts. Specifically, a contact is predicted to be biological if and only if it covers more than $8 \times P_{Cons} + 19$ residues.

The heuristic predictor for relative assessment, H_{rel}, uses a hierarchical scheme to choose the most likely biological contact among a set of contacts. First, a subset of contacts is defined. Each contact in this subset must cover at least 75% as many residues as the largest contact. From this subset, the contact with the smallest P_{Cons}, ie, the most conserved, is then predicted to be biological.

The heuristic predictors, having no explicit training element, were not cross-validated.

4.3.3.2 Predictor performance: Absolute assessment

We applied more than 20 different neural network predictors to the absolute assessment. These spanned a range of single layer perceptron (SLP, a linear network) and multilayer perceptron (MLP*x*, a nonlinear network with *x* hidden units) architectures. We tested all combinations of the three absolute measures listed in Table 4.1.

Figure 4.8 shows the accuracy of predictors in the absolute assessment, whereas Figure 4.9 shows the performance against random, measured by phi (see Methods), for the same experiments. These figures show only a selection of the interesting results, with complex networks omitted if they are outperformed by simpler ones. For instance, we exclude the SLP with two size-related inputs because simpler SLPs with only one size-related input perform at least as well.

All predictors listed gave a correct classification in 87% or more of cases (Figure 4.8). Phi (Figure 4.9) provides a more balanced performance metric. By comparing the observed classification against that expected by random assignment, it accounts for imbalances in the dataset. For example, phi exposes two

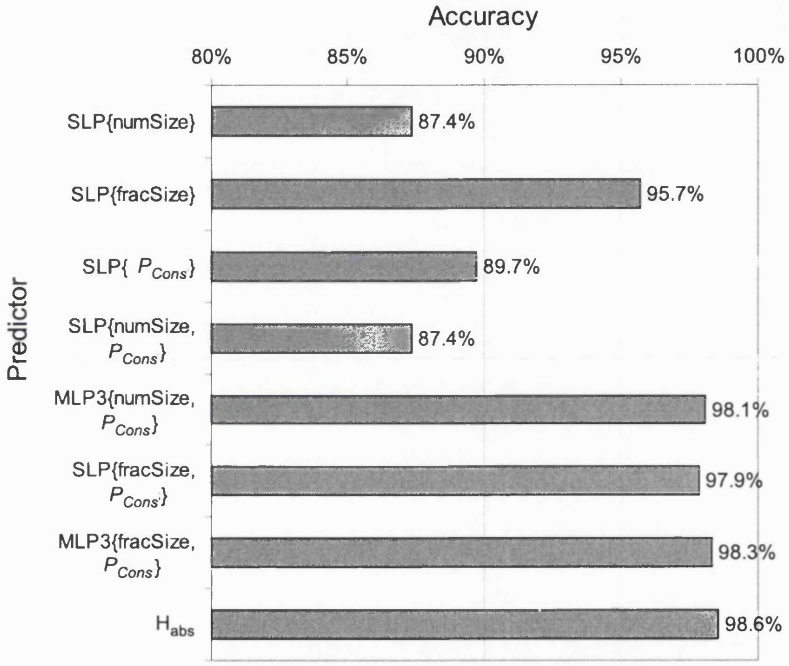


Figure 4.8: Accuracy of predictors in the absolute assessment of homodimer contacts. Predictors are presented on the on the vertical axis. For brevity, neural network predictors are given short names of the form *network{input1, input2, ...}*. The prefix *network* denotes the architecture of the network: SLP (single layer perceptron) or MLP*x* (multilayer perceptron with *x* hidden units). The names in curly braces refer to the inputs of the network: "numSize" is the number of residues in a contact; "fracSize" is the fraction of the surface covered by a contact; P_{Cons} is a measure of the conservation for the contact (see Table 4.1 for fuller definitions). H_{abs} is the heuristic predictor for the absolute assessment and is defined in the Results.. Accuracy, on the horizontal axis, measures the percentage of contacts a predictor correctly classified (see Methods). Assessment of neural network performance is cross-validated with respect to the homodimer dataset. The graph shows that although size alone and conservation alone have predictive power, combining both measures makes predictions more accurate.

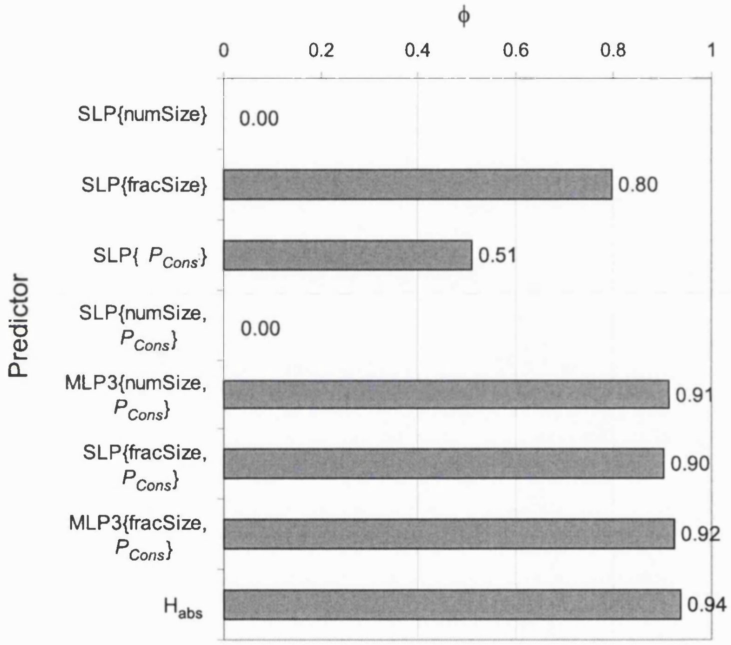


Figure 4.9: Performance of predictors against random in the absolute assessment of homodimer contacts. For an explanation of predictors (vertical axis), see Figure 4.8 legend. Performance is measured by phi (ϕ; see Methods) along the x-axis. Phi is more informative, if less intuitive, than accuracy because it normalizes against imbalances in the dataset. Phi ranges from -1, denoting extremely poor prediction, to 0, denoting prediction equivalent to random assignment, through to +1, denoting a perfect prediction. Assessment of neural network performance is cross-validated with respect to the homodimer dataset. The graph shows linear networks using numSize perform no better than random.

SLPs as poor performers: the SLP with the number of residues in a contact as its sole input, and the SLP that takes both this input and P_{Cons}. Both predictors achieve their apparently high accuracy by predicting all contacts as nonbiological and the imbalanced nature of the dataset means they are correct most of the time. However, as phi confirms, this tactic requires no discriminatory power.

Although conservation alone performed well and size alone, measured as the fraction of surface residues in a contact, performed better, Figures 4.8 and 4.9 show that a combination of size and conservation can be most powerful. For the inputs P_{Cons} and size as a fraction of the surface, the linear network performs significantly better than any network that takes only one of the two. Adding three hidden units improves performance further, although a smaller or greater supplement of hidden units does not. The input combination of P_{Cons} and the number of residues in a contact achieves performance better than random only for MLPs, with architecture MLP3 being optimal.

The best neural network predictor in this category was MLP3 with inputs of P_{Cons} and size as a fraction of the surface. This correctly classified 48 out of the 53 biological contacts and 364 of the 366 nonbiological contacts, giving a combined accuracy of $(48+364)/(53+366) = 98.3\%$. However, even this was outperformed by the heuristic predictor $\mathrm{H_{abs}}$, which relies on nothing more than a linear separation.

Figure 4.10 shows the performance of the same set of predictors in the absolute assessment of crystal contacts in monomers. Because there are no biological contacts in this set, the performance of predictors is ineligible for phi and most meaningfully interpreted with a pure error rate. Again, the seemingly perfect performances of the SLP with the number of contact residues as a single input and the SLP with this and P_{Cons} as dual inputs are specious and owe nothing to discriminatory power. The error rates show SLP with P_{Cons} as sole input, MLP3 with dual inputs of P_{Cons} and the number of contact residues, and $\mathrm{H_{abs}}$ are most prone to overpredict biological interfaces. In contrast, the SLP with the sole input of size as a fraction of the surface, and the SLP and MLP3 with both that size input and P_{Cons} are most discriminating in this respect.

4.3.3.3 Predictor performance: Relative assessment

We tracked the performance of more than 40 different neural network predictors in the relative assessment. The networks tested ranged from those with a single input from Table 4.1 to all six inputs, and some MLPs had as many as four hidden units.

Figure 4.11 shows classification accuracy for a representative selection of the predictors tested. Predictors using relative measures (ie, ranked size, size difference from largest contact or ranked P_{Cons}) as inputs typically performed better than those relying on only absolute measures (ie, number of residues in a contact, contact size as fraction of the surface, and P_{Cons}) (see table 4.1 for an explanation of how these inputs are defined).

Predictors relying on one of ranked size or the size difference from the largest contact achieved correct classifications for all but one protomer, 1uby, and attained the maximum accuracy achieved for any neural network at 98%. 1uby, as mentioned above, is the only protomer for which the biological contact is not the largest crystal contact made. Thus, 98% also corresponds to the predictive accuracy associated with simply designating as biological the largest observed contact. Alone, the relative measure of ranked P_{Cons} had some predictive power (68%). Combining ranked P_{Cons} with ranked size in a linear network gave performance intermediate between that of ranked P_{Cons} alone and ranked size alone. Successive addition of hidden units ameliorated performance, with MLP3 being maximal (data not shown). No such gradient of improvement was seen when combining ranked P_{Cons} with size difference from the largest contact; even with a linear network, performance equaled the maximal 98%. Neural networks relying on only absolute measures for inputs showed the same performance relative to each other in the relative assessment as they did in the

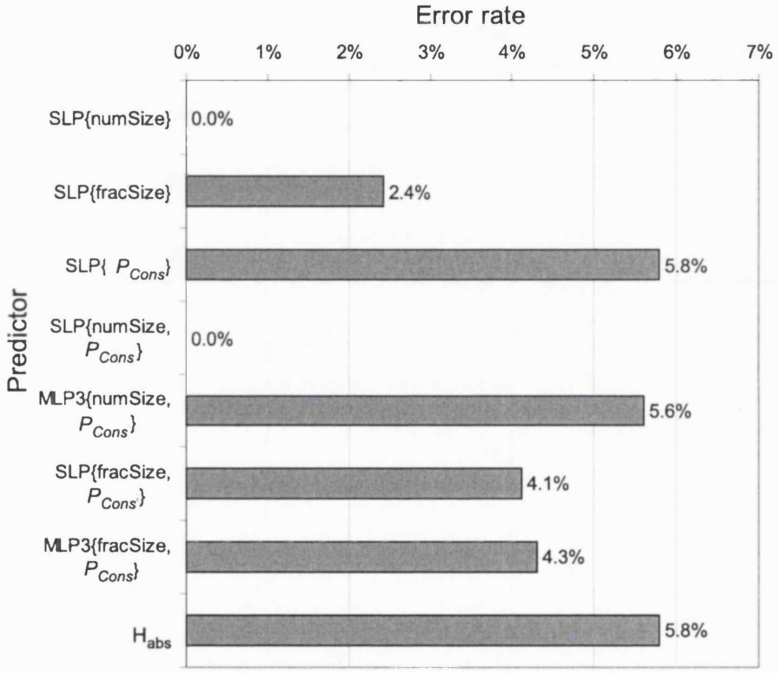


Figure 4.10: Error rate of predictors in the absolute assessment of monomer contacts.

For an explanation of predictors (vertical axis), see Figure 4.8 legend. Error rate (horizontal axis) refers to the percentage of contacts misclassified by a predictor. In this case, the error rate corresponds to the rate of overprediction of biological contacts. All neural networks depicted have been trained on the full homodimer set. The graph shows that the heuristic predictor H_{abs} and the single layer perceptron with only conservation as its input most often overpredict biological contacts.

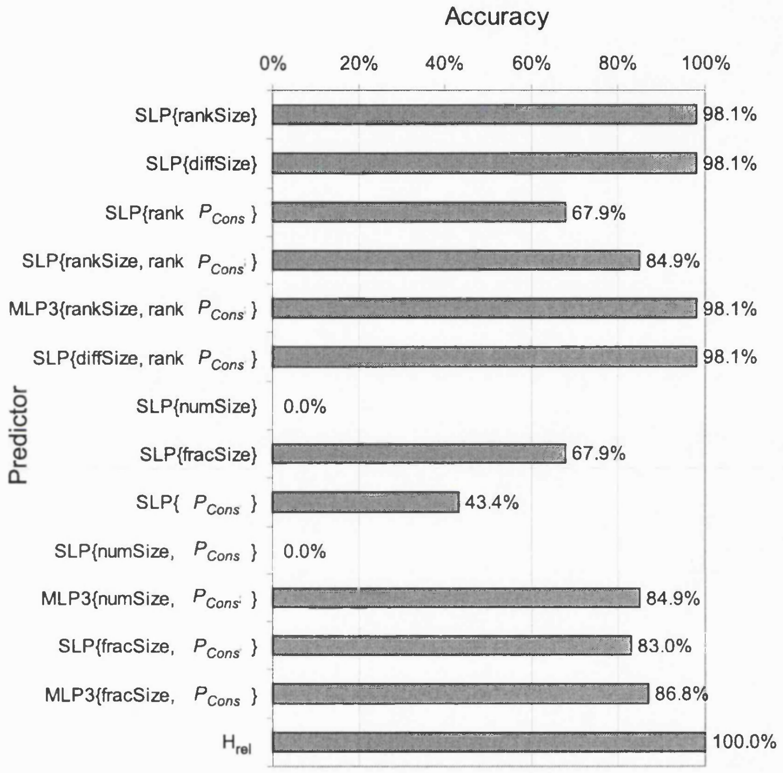


Figure 4.11: Accuracy of predictors in the relative assessment of homodimers.
Predictors are presented on the on the vertical axis. As in earlier figures, neural network predictors are given a short names of the form *network{input1, input2, ...}*, where the prefix *network* denotes the architecture of the network: SLP (single layer perceptron) or MLPx (multilayer perceptron with x hidden units). The names in curly braces refer to the inputs of the network: "numSize" is the number of residues in a contact; "fracSize" is the fraction of the surface covered by a contact; P_{Cons} is a measure of the conservation for the contact; "rank P_{Cons}" is the ranked conservation of a contact, "rankSize" is the ranked size of a contact, and "diffSize" is the difference in size between a given contact and the largest contact in the protomer. The inputs P_{Cons}, numSize and fracSize are "absolute" measures whereas the inputs rank P_{Cons}, rankSize and diffSize are "relative" measures (see Table 4.1 for fuller definitions). H_{rel} is the heuristic predictor for the relative assessment and is defined in the Results. Accuracy, presented on the horizontal axis, measures the percentage of protomers for which a predictor unambiguously identified the biological contact. Assessment of neural network performance is cross-validated with respect to the homodimer dataset. The graph shows that relative measures have more predictive power than absolute measures, and that size alone allows near-perfect prediction.

absolute assessment.

The heuristic method H_{rel} beat all the neural network predictors, correctly predicting the biological contact in every case. But the difference between the 98% of choosing the biggest and 100% for H_{rel} represents more than simply dealing with one recalcitrant protomer. For six of the 53 protomers, H_{rel} used conservation to choose which contact was biological.

4.4 Discussion

The results show biological crystal contacts are typically larger and more conserved than nonbiological ones. Our analysis of contact size agrees with that of previous studies. It finds biological contacts are invariably large and usually the largest contact made in a crystal. Figure 4.1 suggests there may be some upper bound on the size of nonbiological contacts. The reason for this could be principally biophysical. Consider two interacting protomer surfaces. Small sites of interaction that are flat or complementary, either in a geometric or an electrostatic sense, will be common. But, because most protomers are globular, larger sites with these properties will be comparatively rare. Thus, unless there has been evolutionary pressure making a large site advantageous, the probability of a substantial interaction will be low. Another interpretation is that protomers are under selective pressure to avoid forming large interfaces at random. After all, if large random interactions were to occur, this would typically encumber a protomer's function.

The biological contact is the most conserved of all contacts in the crystal much more frequently than would be expected for a random distribution ($P_{mostcon} = 2.38 \times 10^{-19}$). Moreover, the results of the association tests in Table 4.4 show that the presence of nearby conserved ligand-binding sites is not the chief source of this high conservation.

Where biological contacts are not conserved, two kinds of explanation usually prevail. The first relates to the availability of sequence data. Too little variation in the multiple alignment, caused by either too few sequences or too little diversity among them, can render analysis of conservation meaningless. The diversity filter *Doss* (described in Methods) goes some way to remedy this, but the example of 1tox suggests its cutoff could be stricter. Fragmented sequences in the databases can also distort the evolutionary information an alignment provides. If the residues of the biological interface are concentrated around sequence termini, which are often missing in fragments, this problem should be considered (by inspection of component sequences). The second kind of explanation is biological. Different dimerization modes (1alo) and multiple multimeric states in a homologous family (1ad3, 1aor, 1slt and 2tct) help explain why some biological contacts will not be conserved. Given that most aligned sequences lack annotation, it seems likely that the prevalence of these phenomena is underpredicted.

Nonbiological contacts were usually poorly conserved, regardless of whether biological contacts were present in the same crystal (Figure 4.3). A rough analysis of ligand binding sites (Table 4.4) suggested the high conservation of some nonbiological contacts owed much to the strict conservation of nearby catalytic or cofactor sites.

The results from the absolute assessment show that whereas size and conservation may be independently useful for classifying contacts as biological or nonbiological, combining these two orthogonal measures provides predictive accuracy greater than either one alone.

The success of multilayer perceptron neural network architectures over single layer ones suggests that a single straight line is not the best predictor. Rather, it implies the optimal decision boundary is something more sophisticated, such as a number of straight lines or curves. However, the success of the heuristic H_{abs}, which is nothing more than a linear discriminant function, belies this suggestion. Two alternative

Protomer	Multimeric state (M/H)[a]	Correct (-/+)[b] Most accurate neural network[c]	H_{abs}
1auo	D	+	-
1cp2	D	+	+
1slt	D	+	-
1xso	D	-	-
1ako	M	+	+
1avp	M	-	-
1feh	M	+	-
1ton	M	+	+

Table 4.5: Performance of the best predictors from the absolute assessment on the protomers misclassified by Ponstingl et al.

[a] M=monomer, D=dimer

[b] A classification is correct (+) only if all contacts in a crystal structure are correctly classified.

[c] The most accurate neural network was the multilayer perceptron with three hidden units that took the two inputs: P_{Cons} and the fraction of the surface residues covered by a contact.

explanations may account for the disagreement. First, H_{abs} was concocted with knowledge of the entire dataset whereas its neural network equivalent was privy to all data except that on which it was to make a prediction. This may have given H_{abs} an unfair advantage. Second, and more plausible, is a training failure on the part of the neural network. Causes of such failure abound and include the network becoming trapped in local minima of the error surface; the error function being inappropriately or unfortunately defined, eg, it does not sufficiently penalize false negative predictions; one of the input weights has become saturated beyond the point it can be usefully modified; and so on (Shepherd, 1997). However, the training and test data are probably too limited to assess reliably the relative merits of linear over non-linear discriminants. Whereas H_{abs} achieved the best phi for homodimers, it may be that the MLP3 architecture with inputs of P_{Cons} and size as a fraction of the surface is more generally applicable, a postulation supported by its superior error rate in the monomer assessment (Figure 4.10). Also, if further parameters not included here were added, eg, physical measures (Jones & Thornton, 1997b) or pair-scores (Ponstingl et al, 2000), a neural network would provide a more robust framework for mining these higher dimensional data than would the heuristic approach.

The performance of predictors in the relative assessment may be usefully compared with the statistical potential of Ponstingl et al (Ponstingl et al, 2000). Ponstingl et al applied statistical potentials to the same dataset to predict whether a given protomer was a homodimer or monomer. Their pair-score misclassified twelve protomers: seven homodimers as monomers and five monomers as homodimers. We found sufficient sequence information for eight of the twelve. Table 4.5 lists these eight protomers and reports how the MLP3 with inputs of P_{Cons} and size as a fraction of the surface, ie, the best performing neural network, and H_{abs} classified their contacts. It shows the MLP3 correctly classified all contacts, thereby also correctly determining multimeric state, in six of these protomers, whereas the heuristic predictor classified all contacts correctly in only three. This interpretation of the results suggests the consolidating power of the neural network may offer advantages over other methods.

The results from the relative assessment show size is an extremely powerful predictor when it comes to singling out the biological contact from a group of contacts. So powerful, in fact, that adding information about residue conservation produces little benefit if any. The results also show information about other contacts in the same crystal can be more useful than absolute measures for this type of assessment. The use of P_{Cons} and fractional size are deliberate attempts to extend the notion of residue conservation and contact size beyond the scope of a single protomer. Conversely, ranked P_{Cons}, ranked size and difference in size from the largest contact represent attempts to do the opposite. Unsurprisingly then, the relative scores are better suited to the relative assessment, and the absolute scores better suited to the absolute assessment.

The heuristic predictor H_{rel} predicted the biological interface with 100% accuracy. It used a hierarchical prediction scheme, first choosing potential biological contacts by their size, then using conservation to break ties when size delivers ambiguous or plural results. Given the observed power of size alone, this seems a more sensible way to use conservation. However, the neural network architectures used here are not well adapted to make their decisions in this way. A given network can only look at the information about one contact at a time. Although the use of ranked P_{Cons}, ranked size and size difference from the largest contact provide some context for a contact, they are a poor second to seeing all of the data at once.

As for the absolute measures, a larger dataset or higher dimensional data may more sharply resolve the relative strengths and weakness of the neural network paradigm versus the hierarchical approach of H_{rel}.

After the majority of this work had been completed, Elcock & McCammon published a paper on a related topic (Elcock & McCammon, 2001). They used conservation to help distinguish between homodimeric and monomeric crystal structures. This is a different question from either of those posed here. They applied their method to the dataset of Ponstingl et al (Ponstingl et al, 2000) and to a large number of proteins in the PQS (Henrick & Thornton, 1998) database. Because they considered only the largest contact in a crystal, their results, which were promising, are not directly comparable with ours. Here we have tried to be more statistically rigorous in assessing the value of conservation in determining the biological relevance of a contact.

4.5 Conclusion

Conservation alone provides information, which is orthogonal to that of size, that is powerful to help predict the biological relevance of a crystal contact. Conservation and size provide a potent combination for discriminating biological from nonbiological contacts. Ultimately, size remains the most powerful discriminator, but conservation can discriminate between borderline cases.

Neural networks generalize the information from homodimer data well, using it to correctly infer biological relevance in the vast majority of monomer contacts. In hindsight, these two measures could be combined in a simple linear manner to produce a powerful predictor. However, it remains to be seen whether the linear separability observed here holds with a larger dataset.

One natural next step is to apply these networks to higher order oligomers. Another is to present the predictors with more input data, such as pair-potentials or physical measures, to further improve their accuracy. A third is to apply the principles demonstrated in this work to the prediction of putative interfaces in heterodimers or transient multimers. For these types of complexes, it is less likely that the most important biological contacts will be seen in the crystal. The challenge then would be to identify potential interaction surfaces and then screen them using the criteria applied in this paper.

In some oligomers it is clear why a multimeric state is important for their function. In others, the advantage conferred is not obvious. It is particularly interesting to investigate biological contacts that are unconserved. These often reflect the existence of multiple multimeric states, which in turn can be interpreted in two ways. It either shows the contact has no biological importance and therefore has been under no selective pressure to be conserved, or reflects the specialization of different members of the family to perform different functions.

In distinguishing biological from nonbiological crystal contacts, some categories of proteins are more difficult than others. A protomer that participates in multiple interactions, such as a signalling protein or a highly regulated enzyme, may have multiple functional surfaces. In this case, although the biological partner may not be present in the crystal, the corresponding functional interface may provide a non-natural crystal contact that appears conserved. Conservation analysis is therefore useful even when the function of a protein is unknown in that it can identify functional residues. However, when function is known it can help to elucidate the molecular mechanism of biological function and provide clues to be tested experimentally.

Chapter 5

Conclusions

This work has sought to answer three questions. First, can residue conservation be quantified? Second, are protein-protein interfaces conserved? Third, can the conservation of protein-protein interfaces be useful in their prediction? All three questions have been answered in the positive.

Biologically important interfaces are under functional constraints, which in a competitive environment translate into selective pressure. Selective pressure usually manifests itself in a protein family as conservation, but not always. After gene duplication, selective pressure can foster innovation, which leads to diversity. Multiple specificities and multiple multimeric states within protein families, as found in chapter 4, are examples of diversity that is advantageous to the host organism. The residues that dictate which multimeric state a homologue has or which small molecule a homologue binds are among the most functionally important residues in a protein. But paradoxically, they may also be the least conserved. Any analysis of conservation is therefore subject to the following uncertainty: is the position X unconserved because it is unimportant for function or because it dictates function? One way to resolve such uncertainty is to examine the literature and ask if the changes in amino acid at a set of positions correlate with small changes in function. Another is to make an assumption about nature's parsimony. Close homologues, those that have sequence identity of >40%, are likely to have identical functions and have been under identical selective pressure since diverging from a common ancestor. Variant residues will tend to represent susceptibility to genetic drift rather than biological innovation. It can thus be confidently inferred such positions are unimportant for function. More distant homologues are more likely to have subtly different functions or perhaps the same function optimized for a subtly different environment. The significance of positional variability among these proteins is less certain, and in this case a literature review may be warranted. So, when it comes to choosing homologues for analysis of conservation, is closer better? No, because the set of positions that are truly conserved, ie, the conserved signal, is often drowned out by false conservation from positions that have not had time to diverge, ie, conserved noise. To remedy this, a conservation score such as C_{Valdar} downweights the contribution of highly similar sequences. In doing so it also upweights distant homologues, which certainly improves the signal to noise ratio for conservation but may confuse the analysis with genuine biological diversity. In this work, we assume genuine biological diversity in our alignments is the exception rather than the rule. This assumption makes the analysis of conservation tractable and, as shown by the utility of this neutralist position in chapter 4, provides the right answers most of the time.

The score C_{Valdar} was proposed to measure conservation. If conservation is isomorphic with functional constraint, then the success of using C_{Valdar} to discriminate biological from nonbiological contacts in chapter 4 shows this score performs well. It fails sometimes, but most of these cases can be explained in terms of advantageous residue diversity, which is where we would expect it to fail.

Chapter 3 showed that protein-protein interfaces are conserved to a significant degree. That observation owes much to the probabilistic scheme used to compare sets of surface residues. Using raw values of residue conservation, small groups of highly conserved residues that form ligand-binding sites would have seemed the most conserved. P values allowed the comparison of like with like and showed that although small clusters with high average conservation are common, large clusters with moderately high conservation are rare. But although the picking P values showed interfaces were conserved, many of the walking P values were equivocal and suggested their use in full-scale prediction, ie, locating a interface given only the protomer structure and a corresponding sequence alignment, might be more limited.

Full-scale prediction of interfaces was not attempted in chapter 4. Instead, discrimination of biological from nonbiological interfaces, the less ambitious cousin of prediction, was. Relatively high conservation was shown to be a consistent feature of biological interfaces; the graph that used Bayes theorem to measure $P(\text{biological}|P_{Cons})$ (figure 4.4) says it all. But although conservation proved powerful as a discriminator, it was no match for contact size. This suggests conservation is not ideally placed as a sole predictor. If its performance in discrimination is less than perfect, then its performance in a full-scale prediction is likely to be far worse. However, chapter 4 showed conservation adds significant value to size. In the absolute assessment of that chapter, the phi-coefficients of neural networks that combined the two measures were significantly higher than the best single-input network. Of course, size is only one of a number of biophysical measures that could be devised to aid discrimination and prediction. For example, hydrophobicity, residue interface propensities and charge complementary may all provide helpful additional inputs. Nevertheless, conservation should always be useful in combination because, being derived from a historical study of inferred selective pressure, it is orthogonal to all of these measures.

There is more to life than homodimers. In the interests of simplicity, the sweep of this study has been narrow. Complexes that are arguably more interesting to the biologist, such as hetero-oligomers or transient complexes, have been ignored. This was unfortunate but, owing to the current paucity of data for these types of complexes, necessary for such a thorough investigation. It could be that residues in these types of complexes are more consistently conserved. Of course, the opposite may also be true. Correlated mutations, which would confuse our neutralist approach, might be more frequent in such complexes. What is likely is that biophysical measures would be less useful. Functionally important interfaces in any kind of complex are axiomatically under strong selective pressure. However, they are not necessarily flat, large or hydrophobic. The analysis of conservation in other types of complex is thus an exciting and potentially fruitful avenue for further research.

Some protein-protein associations are forever whereas others are fleeting trysts. To make this study tractable, binding constants were ignored. It is interesting to consider how much the strength of an association can be related to the conservation of residues that secure it. A mutation that causes a small change in the binding constant of an interaction might be tolerated by the host organism; some enzymes could be up-regulated, the network of interactions could be adjusted. Depending on the importance of the protein and the precise nature of its interaction, this might debilitate the host or leave the host unaffected. For instance, would a slightly decreased affinity between a G-protein coupled receptor, which could be tolerated by up-regulation of GPCR production and the like, be as cataclysmic as a slightly decreased affinity between the tubulin subunits of microtubules, which may not be remedied so easily? Questions like these suggest there is rich scope for the analysis of conservation in a greater variety of complexes than is studied here. As experimental science continues to elucidate and record the biophysical properties of such complexes, the analysis of evolution in these systems can become subtler. This all relates strongly to interface prediction.

The more the interplay between evolutionary conservation and biophysics is understood, the more sensibly measures relating to these orthogonal perspectives can be combined in a predictive scheme.

Bibliography

Abrahams, J. P., Leslie, A. G., Lutter, R., & Walker, J. E. (1994). Structure at 2.8Å resolution of F1-ATPase from bovine heart mitochondria. *Nature,* **370** (6491), 621–628.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1994). *Molecular biology of the cell.* New York: Garland, 3rd edition.

Altschul, S. F., Carroll, R. J., & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **207**, 647–653.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids. Res.* **25**, 3389–3402.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.

Armon, A., Graur, D., & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.

Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N., & Wright, W. (1999). PRINTS prepares for the new millennium. *Nucl. Acids. Res.* **27**, 220–225.

Babbitt, P. C. & Gerlt, J. A. (1997). Understanding enzyme superfamilies: Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30591–30594.

Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R. J., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., & Gerlt, J. A. (1996). The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids. *Biochemistry,* **35**, 16489–16501.

Baczkowski, A. J., Joanes, D. N., & Shamia, G. M. (1997). Properties of a generalized diversity index. *J. Theor. Biol.* **188**, 207–213.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids. Res.* **28**, 45–48.

Banci, L., Benedetto, M., Bertini, I., Del Conte, R., Piccioli, M., & Viezzoli, M. S. (1998). Solution structure of reduced monomeric Q133M2 copper, zinc superoxide dismutase (SOD). why is SOD a dimeric enzyme? *Biochemistry,* **37**, 11780–11791.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids. Res.* **28**, 263–266.

Bell, C. & Eisenberg, D. (1996). Crystal structure of diphtheria toxin bound to nicotinamide adenine dinucleotide. *Biochemistry,* **35** (4), 1137–1149.

Benner, S. A., Cohen, M. A., & Gonnet, G. H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **7** (11), 1323–1332.

Bennett, M. J., Schlunegger, M. P., & Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **4** (12), 2455–2468.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids. Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L. & Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. USA,* **93**, 14243–14248.

Board, P. G., Baker, R. T., Chelvanayagam, G., & Jermiin, L. S. (1997). Zeta, a novel class of glutathione transferases in a range of species from plants to humans. *Biochem. J.* **328**, 929–935.

Board, P. G., Coggan, M., Wilce, M. C. J., & Parker, M. W. (1995). Evidence for an essential serine residue in the active-site of the theta-class glutathione transferases. *Biochem. J.* **311**, 247–250.

Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.

Borchert, T. V., Abagyan, R., Jaenicke, R., & Wierenga, R. K. (1994). Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. *Proc. Natl. Acad. Sci. USA,* **91**, 1515–1518.

Bordo, D., Djinovic, K., & Bolognesi, M. (1994). Conserved patterns in the Cu,Zn superoxide-dismutase family. *J. Mol. Biol.* **238**, 366–386.

Bordo, D., Matak, D., Djinovic-carugo, K., Rosano, C., Pesce, A., Bolognesi, M., Stroppolo, M. E., Falconi, M., Battistoni, A., & Desideri, A. (1999). Evolutionary constraints for dimer formation in prokaryotic Cu,Zn superoxide dismutase. *J. Mol. Biol.* **285**, 283–296.

Branden, C. & Tooze, J. (1998). *Introduction to protein structure.* New York and London: Garland Publishing Inc., second edition.

Bray, J. E., Todd, A. E., Pearl, F. M. G., Thornton, J. M., & Orengo, C. A. (2000). The CATH dictionary of homologous superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.* **13**, 153–165.

Buckle, A. M., Schreiber, G., & Fersht, A. R. (1994). Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-Å resolution. *Biochemistry,* **33** (30), 8878–8889.

Carugo, O. & Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci.* **6**, 2261–2263.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature,* **256**, 705–708.

Clamp, M. E., Cuff, J. A., & Barton, G. J. (1998). Jalview: a java multiple sequence alignment viewer and editor. http://barton.ebi.ac.uk/.

Creighton, T. E. (1996). *Proteins: structures and molecular properties.* New York: W. H. Freeman and Co., 2nd edition.

Crosio, M. P., Janin, J., & Jullien, M. (1992). Crystal packing in six crystal forms of pancreatic ribonuclease. *J. Mol. Biol.* **228** (1), 243–251.

D'Alessio, G. (1999). The evolutionary transition from monomeric to oligomeric proteins: tools, the environment, hypotheses. *Prog. Biophys. Molec. Biol.* **72**, 271–298.

Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. H., & Bell, J. A. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins,* **28**, 494–514.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). *Atlas of protein sequence and structure* volume 5 chapter A model of evolutionary change in proteins: Matrices for detecting distant relationships, pp. 345–358. Washington, D. C.: National biomedical research foundation.

de Bono, E. (1999). *Simplicity.* UK: Penguin Books.

de Rinaldis, M., Ausiello, G., Cesareni, G., & Helmer-Citterich, M. (1998). Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol.* **284**, 1211–1221.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry,* **29** (31), 7133–7155.

Dubose, R. F. & Hartl, D. L. (1990). The molecular evolution of bacterial alkaline phosphatase: Correlating variation among enteric bacteria to experimental manipulations of the protein. *Molecular Biology And Evolution,* **7**, 547–577.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge, UK: Cambridge University Press.

Durbin, S. D. & Feher, G. (1996). Protein crystallization. *Annu. Rev. Phys. Chem.* **47**, 171–204.

Eddy, S. R. (1996). Hidden markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.

Eddy, S. R., Mitchison, G., & Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.* **2**, 9–23.

Elcock, A. H. & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. USA,* **98** (6), 2990–2994.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402** (6757), 86–90.

Enroth, C., Eger, B. T., Okamoto, K., Nishino, T., Nishino, T., & Pai, E. F. (2000). Crystal structures of bovine milk xanthine dehydrogenase and xanthine oxidase: structure-based mechanism of conversion. *Proc. Natl. Acad. Sci. USA,* **97** (20), 10723–10728.

Etzold, T., Ulyanov, A., & Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. *Methods Enzymol.* **266**, 114–128.

Fisher, R. A. (1990). *Statistical methods, experimental design, and scientific inference.* Oxford: Oxford University Press.

Garza-Ramos, G., Cabrera, N., Saavedra-Lira, E., Degomez-Puyou, M. T., Ostoa-Saloma, P., Perez-Montfort, R., & Gomez-Puyou, A. (1998). Sulfhydryl reagent susceptibility in proteins with high sequence similarity – triosephosphate isomerase from *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania mexicana*. *Eur. J. Biochem.* **253**, 684–691.

Göbel, U., Sander, S., Schneider, R., & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins,* **18**, 309–317.

Gerstein, M. & Altman, R. B. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175.

Gerstein, M., Sonnhammer, E. L. L., & Chothia, C. (1994). Volume changes in protein evolution. *J. Mol. Biol.* **236**, 1067–1078.

Gerstein, M., Tsai, J., & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249** (5), 955–966.

Getzoff, E. D., Tainer, J. A., Stempien, M. M., Bell, G. I., & Hallewell, R. A. (1989). Evolution of CuZn superoxide-dismutase and the greek key β-barrel structural motif. *Proteins,* **5**, 322–336.

Gigant, B., Curmi, P. A., Martin-Barbey, C., Charbaut, E., Lachkar, S., Lebeau, L., Siavoshian, S., Sobel, A., & Knossow, M. (2000). The 4Å X-ray structure of a tubulin: stathmin-like domain complex. *Cell,* **102** (6), 809–816.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika,* **40**, 237–264.

Goodsell, D. S. & Olson, A. J. (1993). Soluble proteins: size, shape and function. *Trends Biochem. Sci.* **18** (3), 65–68.

Goodsell, D. S. & Olson, A. J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153.

Gopal, B., Ray, S. S., Gokhale, R. S., Balaram, H., Murthy, M. R. N., & Balaram, P. (1999). Cavity-creating mutation at the dimer interface of *Plasmodium falciparum* triosephosphate isomerase: restoration of stability by disulfide cross-linking of subunits. *Biochemistry,* **38**, 478–486.

Gregory, R. L. & Zangwill, O. L. (1987). *The Oxford companion to the mind.* Oxford, UK: Oxford University Press.

Grishin, N. V. & Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.* **3**, 2455–2458.

Hedstrom, L. (1996). Trypsin: a case study in the structural determinants of enzyme specificity. *Biol. Chem.* **377**, 465–470.

Hendrich, M. P., Petasis, D., Arciero, D. M., & Hooper, A. B. (2001). Correlations of structure and electronic properties from EPR spectroscopy of hydroxylamine oxidoreductase. *J. Am. Chem. Soc.* **123** (13), 2997–3005.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids. Res.* **19**, 6565–6572.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89**, 10915–10919.

Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.

Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.

Henriques, D. A., Ladbury, J. E., & Jackson, R. M. (2000). comparison of binding energies of SrcSH2-phosphotyrosyl peptides with structure-based prediction using surface area based empirical parameterization. *Protein Sci.* **9**, 1975–1985.

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383–402.

Hill, R. E. & Hastie, N. D. (1987). Accelerated evolution in the reactive center regions of serine protease inhibitors. *Nature,* **326**, 96–99.

Hille, R. (1999). Molybdenum enzymes. *Essays in biochemistry,* **34**, 125–137.

Hirono, S., Akagawa, H., Mitsui, Y., & Iitaka, Y. (1984). Crystal structure at 2.6 Ångstrom resolution of the complex of subtilisin BPN' with streptomyces subtilisin inhibitor. *J. Mol. Biol.* **178**, 389–413.

Hubbard, S. J. & Thornton, J. M. (1993). NACCESS [computer program]. Department of biochemistry and molecular biology, University College London.

Hulett, F. M., Kim, E. E., Bookstein, C., Kapp, N. V., Edwards, C. W., & Wyckoff, H. W. (1991). Bacillus-subtilis alkaline phosphatase iii and phosphatase iv: Cloning, sequencing, and comparisons of deduced amino acid sequence with escherichia coli alkaline phosphatase three-dimensional structure. *J. Biol. Chem.* **266**, 1077–1084.

Huynen, M., Snel, B., Lathe, W., & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10** (8), 1204–1210.

Hyvönen, M., Begun, J., & Blundell, T. (2000). *Protein-protein recognition* chapter Protein-protein interactions in eukaryotic signal transduction, pp. 189–227. Oxford, UK: Oxford University Press.

Igarashi, N., Moriyama, H., Fujiwara, T., Fukumori, Y., & Tanaka, N. (1997). The 2.8 Å structure of hydroxylamine oxidoreductase from a nitrifying chemoautotrophic bacterium, *Nitrosomonas europaea. Nat. Struct. Biol.,* **4** (4), 276–284.

Iyer, G. H., Dasgupta, S. D., & Bell, J. A. (2000). Ionic strength and intermolecular contacts in protein crystals. *J. Cryst. Growth,* , **217**, 429–440.

Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.,* **4** (12), 973–974.

Janin, J. (2000). *Protein-protein recognition*, chapter Kinetics and thermodynamics of protein-protein interactions, pp. 1–32. Oxford, UK: Oxford University Press.

Janin, J., Miller, S., & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.,* **204**, 155–164.

Janin, J. & Rodier, F. (1995). Protein-protein interactions at crystal contacts. *Proteins,* , **23**, 580–587.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Jones, S. & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Molec. Biol.*, **63**, 31–65.

Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA,* , **93**, 13–20.

Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

Jones, S. & Thornton, J. M. (2000). *Protein-protein recognition*, chapter Analysis and classification of protein-protein interactions from a structural perspective. Oxford, UK: Oxford University Press.

Jores, R., Alzari, P. M., & Meo, T. (1990). Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. *Proc. Natl. Acad. Sci. USA,* , **87**, 9138–9142.

Karlin, S. & Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. *J. Bacteriol.*, **178** (7), 1881–1894.

Kim, E. E. & Wyckoff, H. W. (1991). Reaction mechanism of alkaline phosphatase based on crystal structures: Two-metal ion catalysis. *J. Mol. Biol.*, **218**, 449–464.

Kleanthous, C. & Pommer, A. J. (2000). *Protein-protein recognition*, chapter Nuclease inhibitors, pp. 282–311. Oxford, UK: Oxford University Press.

Knowles, J. R. (1991). Enzyme catalysis: not different, just better. *Nature,* , **350**, 121–124.

Kojima, S., Kumagai, I., & Miura, K. (1993). Requirement for a disulfide bridge near the reactive site of protease inhibitor SSI (*streptomyces* subtilisin inhibitor) for its inhibitory-action. *J. Mol. Biol.*, **230**, 395–399.

Kraulis, P. J. (1991). MolScript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, **24**, 946–950.

Landgraf, R., Fischer, D., & Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.*, **12** (11), 943–951.

Larsen, T. A., Olson, A. J., & Goodsell, D. S. (1997). Morphology of protein-protein interfaces. *Curr. Biol.*, **6**, 421–427.

Larsen, T. M., Wedekind, J. E., Rayment, I., & Reed, G. H. (1996). A carboxylate oxygen of the substrate bridges the magnesium ions at the active site of enolase: structure of the yeast enzyme complexed with the equilibrium mixture of 2- phosphoglycerate and phosphoenolpyruvate at 1.8 Ångstrom resolution. *Biochemistry,* , **35**, 4349–4358.

Laskowski, M. J. R. & Kato, I. (1980). Protein inhibitors of proteinases. *Annu. Rev. Biochem.*, **49**, 593–626.

Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein-protein interfaces. *J. Mol. Biol.*, **234**, 946–950.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Liao, D., Kapadia, G., Ahmed, H., Vasta, G., & Herzberg, O. (1994). Structure of S-lectin, a developmentally regulated vertebrate beta-galactoside-binding protein. *Proc. Natl. Acad. Sci. USA*, , **91** (4), 1428–1432.

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lichtarge, O., Yamamoto, K. R., & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.*, **274** (3), 325–337.

Lijnzaad, P. & Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, , **28**, 333–343.

Liu, Z.-J., Sun, Y.-J., Rose, J., Chung, Y.-J., Hsiao, C.-D., Chang, W.-R., Kuo, I., Perozich, J., Lindahl, R., Hempel, J., & Wang, B.-C. (1997). The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the rossmann fold. *Nat. Struct. Biol.*, **4** (4), 317–326.

Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9** (6), 745–756.

Lo Conte, L., Chothia, C., & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, , **286**, 295–299.

Magalhaes, A., Maigret, B., Hoflack, J., Gomes, J. N., & Scheraga, H. A. (1994). Contribution of unusual arginine-arginine short-range interactions to stabilization and recognition in proteins. *J. Protein Chem.*, **13** (2), 195–215.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, , **285**, 751–753.

Mayer, B. J. (2001). SH3 domains: complexity in moderation. *J. Cell. Sci.*, **114**, 1253–1263.

Mazzarella, L., Capasso, S., Demasi, D., Dilorenzo, G., Mattia, C. A., & Zagari, A. (1993). Bovine seminal ribonuclease - structure at 1.9 Å resolution. *Acta Crystallogr.*, **D49** (4), 389–402.

McCoy, A. J., Epa, V. C., & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.*, **268**, 570–584.

McElroy, H. E., Sisson, G. W., Schoettlin, W. E., Aust, R. M., & Villafranca, J. E. (1992). Studies on engineering crystallizability by mutation of surface residues of human thymidylate synthase. *J. Cryst. Growth*, , **122**, 265–272.

Merritt, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.

Miller, S., Lesk, A. M., Janin, J., & Chothia, C. (1987). The accessible surface-area and stability of oligomeric proteins. *Nature,* , **328**, 834–836.

Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.*, **308**, 123–129.

Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, **291**, 177–196.

Miyata, T., Miyazawa, S., & Yashunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12** (3), 219–236.

Møler, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks,* , **6** (4), 525–533.

Monod, J., Wyman, J., & Changeux, J.-P. (1965). On the nature of allosteric interactions: a plausible model. *J. Mol. Biol.*, **12**, 88–118.

Moont, G., Gabb, H. A., & Sternberg, J. E. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins,* , **35**, 364–373.

Morgan, N. G. (1994). *Cell signalling.* Chichester, UK: John Wiley & Sons.

Mott, R. (2000). Accurate formula for *P*-values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300**, 649–659.

Musacchio, A., Saraste, M., & Wilmanns, M. (1994). High-resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nat. Struct. Biol.* **1** (8), 546–551.

Neuefeind, T., Huber, R., Reinemer, P., Knablein, J., Prade, L., Mann, K., & Bieseler, B. (1997). Cloning, sequencing, crystallization and x-ray structure of glutathione S-transferase-III from zea mays var. mutin: a leading enzyme in detoxification of maize herbicides. *J. Mol. Biol.* **274**, 577–587.

Nogales, E. (2000). Structural insights into microtubule function. *Annu. Rev. Biochem.* **69**, 227–302.

Orengo, C. A. (1999). CORA — topological fingerprints for protein structural families. *Protein Sci.* **8**, 699–715.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). Cath — a hierarchic classification of protein domain structures. *Structure,* **5**, 1093–1108.

Page, R. D. M. & Holmes, E. C. (1998). *Molecular evolution: a phylogenetic approach.* Oxford, UK: Blackwell Science, 2nd edition.

Pazos, P., Helmer-Citterich, M., Ausiello, G., & Valencia, A. (1997). Correlated mutations contain information about protein-protein interactions. *J. Mol. Biol.* **271**, 511–523.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA,* **96**, 4285–4288.

Pilpel, Y. & Lancet, D. (1999). The variable and conserved interfaces of modeled olfactory receptor proteins. *Protein Sci.* **8**, 969–977.

Plaxco, K., Larson, S., Ruczinski, I., Riddle, D., Buchwitz, B., Davidson, A., & Baker, D. (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303–312.

Ponstingl, H., Henrick, K., & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins,* **41**, 47–57.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1996). *Numerical recipes in C: the art of scientific computing*. Cambridge, UK: Cambridge University Press, 2nd edition.

Richardson, D. J. & Watmough, N. J. (1999). Inorganic nitrogen metabolism in bacteria. *Curr. Opin. Chem. Biol.* **3**, 207–219.

Robert, C. H. & Janin, J. (1998). A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J. Mol. Biol.* **283**, 1037–1047.

Romão, M. J., Archer, M., Moura, J. J. G., Legall, J., Engh, R., Schneider, M., Hof, P., & Huber, R. (1995). Crystal-structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*. *Science,* **270**, 1170–1176.

Rossjohn, J., Feil, S. C., Wilce, M. C. J., Sexton, J. L., Spithill, T. W., & Parker, M. W. (1997). Crystallization, structural determination and analysis of a novel parasite vaccine candidate: Fasciola hepatica glutathione s-transferase. *J. Mol. Biol.* **273**, 857–872.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* **9**, 56–68.

Sasso, M. P., Lombardi, M., Confalone, E., Carsana, A., Palmieri, M., & Furia, A. (1999). The differential pattern of tissue-specific expression of ruminant pancreatic type ribonucleases may help to understand the evolutionary history of their genes. *Gene,* **227** (2), 205–212.

Schneider, T. D. (1997). Information content of individual genetic sequences. *J. Theor. Biol.* **189**, 427–441.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M., & Fersht, A. R. (1990). Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry,* **29** (40), 9343–9352.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal,* **27**, 379–423, 623–656.

Sheinerman, F. B., Norel, R., & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10** (2), 153–159.

Shenkin, P. S., Erman, B., & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins,* **11**, 297–313.

Shepherd, A. J. (1997). *Second-order methods for neural networks: fast and reliable training methods for multilayer perceptrons*. London: Springer.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Florida: Chapman & Hall / CRC, 2nd edition.

Shindyalov, I. N., Kolchanov, N. A., & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. *Protein Eng.* **7**, 349–358.

Smith, R. F. & Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. *Protein Eng.* **5** (1), 35–41.

Sternberg, M. J. E., Gabb, H. A., & Jackson, R. M. (1998). Predictive docking of protein-protein and protein-DNA complexes. *Curr. Opin. Struct. Biol.* **8**, 250–256.

Taguchi, S., Kojima, S., Terabe, M., Kumazawa, Y., Kohriyama, H., Suzuki, M., Miura, K., & Momose, H. (1997). Molecular phylogenetic characterization of streptomyces protease inhibitor family. *J. Mol. Evol.* **44**, 542–551.

Takeuchi, Y., Nonaka, T., Nakamura, K. T., Kojima, S., Miura, K., & Mitsui, Y. (1992). Crystal-structure of an engineered subtilisin inhibitor complexes with bovine trypsin. *Proc. Natl. Acad. Sci. USA,* **89**, 4407–4411.

Tamura, A., Kojima, S., Miura, K. I., & Sturtevant, J. M. (1995). A thermodynamic study of mutant forms of streptomyces subtilisin inhibitor. II. replacements at the interface of dimer formation, val13. *J. Mol. Biol.* **249**, 636–645.

Taylor, W. R. (1986). The classification of amino acid conservation. *J. Theor. Biol.* **119**, 205–218.

Taylor, W. R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348.

Teichmann, S. A., Murzin, A. G., & Chothia, C. (2001). Determination of protein fuction, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **11**, 354–363.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., & Higgins, D. G. (1997). The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids. Res.* **25** (24), 4876–4882.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.* **10**, 19–29.

Valdar, W. S. J. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins,* **42** (1), 108–124.

Vijayan, M. & Chandra, N. (1999). Lectins. *Curr. Opin. Struct. Biol.* **9**, 707–714.

Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biosci.* **5**, 115–121.

Williams, J. C., Zeelen, J. P., Neubauer, G., Vriend, G., Backmann, J., Michels, P. A. M., Lambeir, A. M., & Wierenga, R. K. (1999). Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Eng.* **12**, 243–250.

Williamson, R. M. (1995). Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J. Theor. Biol.* **174**, 179–188.

Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci. USA,* **93**, 14249–14255.

Wu, C. H. & McLarty, J. W. (2000). *Neural networks and genome informatics.* Methods in computational biology and biochemistry. Oxford, UK: Elsevier Science Ltd.

Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132**, 211–249.

Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717–729.

Zhang, E., Brewer, J. M., Minor, W., Carreira, L. A., & Lebioda, L. (1997). Mechanism of enolase: the crystal structure of asymmetric dimer enolase-2-phospho-d-glycerate/enolase-phosphoenolpyruvate at 2.0 Ångstrom resolution. *Biochemistry,* **36**, 12526–12534.

Zvelibil, M. J., Barton, G. J., Taylor, W. R., & Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195** (4), 957–961.

Appendix A

Publications arising from this work

The following information was correct as of November 2001:

1. Valdar WSJ (expected 2002) Scoring residue conservation (*submitted*).
2. Valdar WSJ & Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Genetics.* 42(1): 108-124.
3. Valdar WSJ & Thornton JM (2001) Conservation helps to identify biologically relevant crystal contacts. *Journal of Molecular Biology.* 313(2): 399-416.