Genomic fold assignment and rational modeling of proteins of biological interest

J. Michael Sauder and Roland L. Dunbrack, Jr.

Institute for Cancer Research, Fox Chase Cancer Center
7701 Burholme Ave., Philadelphia PA 19111
M_Sauder@fccc.edu and RL_Dunbrack@fccc.edu
phone: 215-728-2434; FAX: 215-728-2412

Abstract

The first available genome of a multicellular organism, *C. elegans*, was used as a test case for protein fold assignment using PSI-BLAST, followed by rational structure modeling and interpretation of experimental mutagenesis data in the context of collaboration with biologists. Similar results are demonstrated for human disease proteins with known polymorphisms.

Introduction

The availability of entire genomic sequences in recent years has made it possible to compare the genomes of different organisms, as well as evaluate the distribution of known protein structures expressed by a particular organism (for examples, see Gerstein, 1998; Wolf *et al.*, 1999). The number of sequenced genomes will most certainly expand rapidly, as will the number of sequenced human genomes, the first of which will be available this year. The expected redundancy of genomic data for a given species (esp. *Homo sapiens*) will also allow wide-scale classification of polymorphisms, or natural amino acid variants in certain proteins.

Rapid methods have been developed and successfully applied for identifying genomic proteins related to proteins of known function from any organism. The most common detection methods use profiles created by PSI-BLAST (Altschul *et al.*, 1997; Altschul *et al.*, 1998) or hidden Markov models (Karplus *et al.*, 1998). These profiles can be costly to create, in terms of computer time, but can very rapidly search a genome containing thousands of sequences.

When the parent (or template) sequence used to build the profile is that of a protein of known structure from the Protein Databank (PDB) (Berman *et al.*, 2000), it is often possible to build a three-dimensional model of the genomic (or target) protein, whose structure is likely unknown. This homology, or comparative, modeling can be done for a significant fraction of genomic proteins; currently, portions of 20-40% of genomic proteins can be modeled, representing from 15% to 35% of the genome on a per-residue basis (unpublished data). This percentage will grow with the number of novel structures deposited in the PDB. Recent structural genomics initiatives (Burley *et al.*, 1999), intended to solve the structures of novel folds, will bolster this growth, but comparative modeling will still be necessary to fill the ~50-fold gap between the number of experimentally determined sequences and structures.

Although several research groups have benchmarked the detection capability of BLAST and PSI-BLAST (Brenner *et al.*, 1998; Park *et al.*, 1998), no groups have performed a rigorous analysis of the accuracy of the alignments generated by BLAST/PSI-BLAST. Since the quality of an alignment is the single most important factor in homology modeling, we also assessed the alignment accuracy of BLAST and PSI-BLAST as a precursor to genomic fold assignment (Sauder *et al.*, 2000).

However, a protein structure model based on a homologous PDB structure is of little or no value unless there is a context for its use and interpretation. The utility of a model is generally correlated to the amount of experimental data that is available for the protein and/or how much is already known about the structure and function of the template (parent) protein. Databases such as ModBase (Sanchez *et al.*, 2000) and Swiss-Model (Guex *et al.*, 1997) provide large numbers of automatically generated models, but the value of homology models is only realized when they are interpreted in light of experimental data.

In an attempt to perform rational structure modeling, we chose to build models of *C. elegans* proteins that are under active investigation by biologists. This inventory of proteins was assembled based on talks and posters at the 12th International *C. elegans* Meeting (June 1999), as well as from available entries in the WormPD at http://www.proteome.com/databases/WormPD (Costanzo *et al.*, 2000) , a *C. elegans* protein database created through exhaustive distillation of the worm literature. This approach gave us access to experimental data, such as mutant genotypes and phenotypes, as well as the ability to correspond with biologists and attempt to answer specific, relevant questions through interpretation of the structure models.

In the case of the human genome, we are interested in proteins for which polymorphisms have been linked to susceptibility or resistance to disease. Polymorphisms are natural amino acid variations in a given protein within a population; polymorphic data are increasingly available in

such databases as HGMD (Krawczak *et al*., 2000) and OMIM (Hamosh *et al*., 2000). Two examples are given for models of human disease proteins where polymorphic data was available.

Methods

Fold assignment

Fold assignment was performed on the most recent version of the *C. elegans* wormpep database, currently wormpep18 containing 18,576 sequences, which is available at ftp.sanger.ac.uk/pub/databases/wormpep/. PSI-BLAST profiles based on sequences of known structure were created using the Astral/SCOP domain database (Brenner *et al*., 2000), currently version 1.48. The 4,466 Astral sequences shared less than 95% identity and domains with non-consecutive sequence regions were excluded. The sequences were modified so that non-standard amino acids (represented as X) were replaced by the most closely related standard amino acid (For example, selenomethionines are Met, phosphorylated tyrosines are Tyr, etc. The S2C database (http://www.fccc.edu/research/labs/dunbrack/s2c/) provides these sequences and correlates SEQRES and ATOM numbering (Arthur *et al*., 2000)). Profiles were created for each Astral domain sequence by iteratively searching *nrhc* (a non-redundant version of Genbank filtered to mask low-complexity regions) with PSI-BLAST. At most 4 iterations were performed (-j 4) with an E-value cutoff (-h) of 0.0001 for inclusion of sequences into the position-specific matrix. The gap trigger parameter (-N 18) and the threshold for extending hits (-f 8) were lowered to optimize hit detection. Low-complexity filtering of the query sequence (-F T) was used during the profile creation step. Profiles that became polluted after 4 rounds of PSI-BLAST were discarded then repeated with a single iteration. (Pollution was obvious if the top scoring hit in the first round was completely missing in the fourth round.) For sequences that converge after the first round (and no profile is created), checkpoint files were generated manually using the –B option in Blast version 2.0.10. This allowed us to create a library of PSI-BLAST checkpoint files that represented all of the SCOP sequences.

The PSI-BLAST profiles (checkpoint files) were then used to search the wormpep database of *C. elegans* sequences. The data allowed us to tabulate (1) all the *C. elegans* proteins homologous to a particular structural domain or protein family, and (2) all the domains in each *C. elegans* protein for which structural information is known. The extent of coverage was calculated by counting the number of residues in each sequence that were aligned with a SCOP sequence and had an E-value below 0.01, which is a reasonable cutoff using the profile searching method, as long as there is not significant amino acid compositional bias. The SCOP profiles were also compared against a *C. elegans* sequence database containing reversed sequences in order to identify profiles that detect a large number of false positives, and to identify sequences with compositional bias.

The same analysis was performed on human proteins. Over 75,000 sequences were downloaded from GenBank and a non-redundant database of 46,876 sequences was created for use with PSI-BLAST. This does not reflect the number of independent human genes in GenBank, since some genes are represented multiple times because of mutations and alternate splicing.

Model building

The PSI-BLAST alignment between the target (query) *C. elegans* or human sequence and the parent (hit) PDB sequence was used as the basis for homology modeling. A Perl program, blast2model, was created which builds backbone model(s) of one or more proteins from a PSI-BLAST alignment file (similar to the procedure in Dunbrack, 1999). In addition, (1) a RasMol (Sayle *et al*., 1995) script file is created which highlights conserved and variant sidechains and maps the gaps in the alignment onto the structure, and (2) an input file is created for building variant sidechains on the model using SCWRL (Bower *et al*., 1997; Dunbrack *et al*., 1997), which predicts the most probable sidechain conformations using a backbone-dependent rotamer library. Blast2model is available from the authors and is part of the SCWRL distribution.

In positions of the alignment where residues were conserved, the sidechain coordinates from the parent PDB file were kept unchanged in the model. If the residues differed or the coordinates were incomplete, sidechains were built using SCWRL 2.1. Because the sequence identities are on average less than 25%, no attempt was made to model missing loop regions, which are due to gaps in the alignment. No refinement procedures were used (other than SCWRL), since the current opinion from the CASP experiments suggests that "energy minimization or molecular dynamics generally leads to a model that is less like the experimental structure" (Koehl *et al*., 1999).

Protein selection

Proteins were chosen for modeling based on two or more of the following criteria:

1. The protein showed some homology to a protein of known structure (i.e., it was detected by PSI-BLAST with an E-value smaller than about 1×10^{-10}).
2. The protein was the subject of investigation at the 12th International *C. elegans* Meeting in June 1999 based on talks and posters.
3. Human targets were chosen in cases where there was polymorphism data and a link to disease available.
4. A model was personally requested by a biologist.

Mutation genotype and phenotype

Mutation information was obtained for many *C. elegans* proteins from the WormPD, available at www.proteome.-com. Human gene mutations were obtained from HGMD

(Krawczak *et al*., 2000) and OMIM (Hamosh *et al*, 2000). Other information was obtained directly from biologists. The effect of a mutation on structure, stability and/or function were postulated by identifying its location in the structure in relation to other conserved or critical residues, confirming allowable rotamers based on backbone ϕ,ψ angles and, in some cases, SCWRL was used to model the coordinates of the mutated sidechain to probe the electrostatic consequences of the mutation.

Benchmarking alignment quality

Since our models are based on PSI-BLAST sequence alignments between the target protein and a protein of known structure, we created a structure-based benchmark to assess the accuracy of PSI-BLAST alignments. The structures of all proteins within a given SCOP family and superfamily were structurally aligned using the CE program (Shindyalov *et al.*, 1998), and structure-based sequence alignments were generated to be used as the "gold standard." Alignment from CLUSTALW, pairwise BLAST, profile-based PSI-BLAST, and an intermediate sequence search (ISS) implementation of PSI-BLAST were then compared with the structure alignments and two scores were calculated to measure the alignment accuracy. The first score is the fraction of correct residues in the sequence alignment (as judged by the structure alignment); the second score is the fraction of the structure alignment that is correctly aligned in the sequence alignment. The scores shown in Figure 1 were normalized by the number of aligned pairs in each superfamily in each sequence identity range.

Results

Alignment quality benchmark

The all-against-all structural comparison of 2,622 SCOP protein domains demonstrated that PSI-BLAST generates fairly accurate and long alignments, in contrast to global alignment programs such as CLUSTALW, where the alignments have more errors below 30% sequence identity, and pairwise local alignment programs such as BLAST, where the alignments are accurate but prematurely truncated (Sauder *et al.*, 2000). These two situations are illustrated in Figure 1, where the results from the most difficult superfamily-level comparisons are shown. A total of 30,000 superfamily-level alignments were performed, where 99% of the sequences share less than 30% identity. The top panel reports the fraction correct for each alignment, averaged in bins of 5% identity and normalized by the number of proteins in each superfamily. PSI-BLAST (solid circles) performs the best, approaching the theoretical limit at 5-15% identity based on alignments (solid line) from the FSSP database of structural alignments (Holm *et al.*, 1998). BLAST and ISS-E (where the intermediate sequence is chosen according to the best E-value) also perform well, but the global alignment method (pairwise) CLUSTALW does rather poorly at low identity. Above ~35-40% identity in the SCOP family-level comparisons, all methods perform equally well.

The bottom panel in Figure 1 compares the methods according to the fraction of the structure alignment that they are able to align. This score favors long alignments, not just accurate alignments. BLAST performs the worst, because it's alignments are too short. PSI-BLAST and ISS-E perform better, but even their alignments are rather short compared to the FSSP structure-based alignments. Structure-based methods (threading/fold recognition) will obviously be able to *detect* more remote homologues than sequence-based methods at random-like sequence identities (<10%), but for homologues that can be detected by both profile and structure-based methods—virtually all can be detected above 20% identity using profile-based methods (Sauder *et al.*, 2000)—it is not clear whether there is much difference is the accuracy of the alignments. We and other groups (Domingues *et al.*, 2000) are beginning to address this question.

The accuracy of PSI-BLAST gave us confidence in

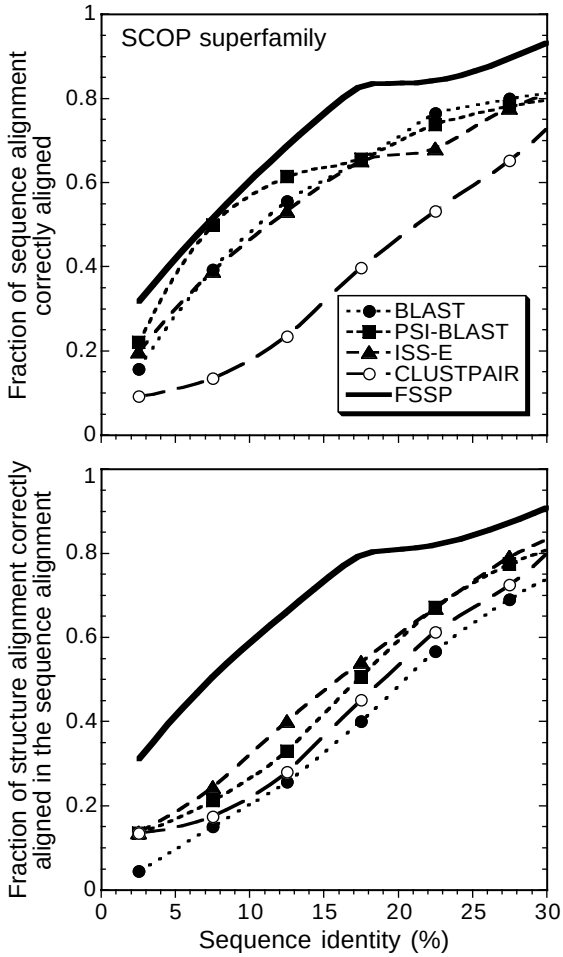

Figure 1. Comparison of the accuracy of 30,000 SCOP superfamily-level alignments using four sequence alignment methods as judged by structure alignments.

using these alignments as the basis for comparative modeling, even if the alignments tend to be shorter than purely structure-based methods. Preliminary results show that PSI-BLAST alignments may be as accurate as those from fold recognition methods, although they are generally shorter.

Genomic fold assignment

Of the 18,576 proteins encoded by the *C. elegans* genome (according to wormpep18), over 34% have some homology to known protein structures, which represents almost 22% of the genome on a per-residue basis (using an E-value cutoff of 0.01). The top 5 folds (such as kinases and nucleotide triphosphate hydrolases) represent almost 10% of the genomic proteins.

Of the 46,876 non-redundant human sequences downloaded from Genbank, 42% had some homology to known structures, with almost 33% of the individual residues aligned with a SCOP domain.

Several examples will be given below of proteins that were modeled, demonstrating what can be learned by integrating fold assignment and rational modeling based on experimental data. The first two examples are both orthologs of mammalian receptor tyrosine kinases (RTKs), although one belongs to class 2 (insulin/IGF1-R) and the other to class 8 (ephrin receptors). The next protein is a serine hydroxymethyltransferase (SHMT) essential to *C. elegans* larval development. The fourth example is a TWIK 4 transmembrane potassium channel protein.

Finally, we provide two examples of the same methodology applied to human sequences with known polymorphisms. The first is the β-secretase involved in cleaving amyloid precursor protein to the amyloid β peptide, responsible for Alzheimer's disease. The second is human cystathionine β-synthase.

DAF-2

DAF-2 is a class 2 insulin-like receptor tyrosine kinase (RTK) (Kimura *et al*., 1997), involved in regulating metabolism, development, fertility and longevity in *C elegans*. Mutations in the protein result in increased life-span, since the worm enters the long-lived dauer stage (DAF = DAuer Formation) instead of entering its reproductive life cycle. It has been hoped that DAF-2 may serve as a link to understanding the relationship between mammalian metabolism and longevity, since insulin-like metabolic control is involved in regulating life-span in the worm (Kimura *et al*., 1997).

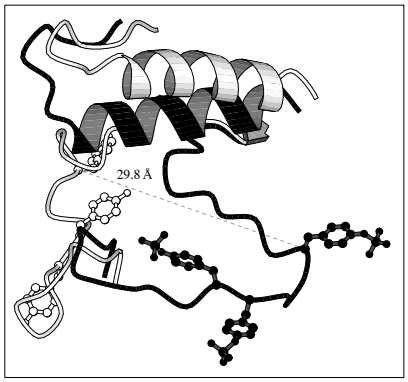

Figure 2. Superposition of the phosphorylated (black) and unphosphorylated (white) form of the kinase domain of the insulin-like RTK (DAF-2), showing movement of the activation loop and the helix near the active site.

The N-terminal receptor domain of DAF-2 shares 32% identity with the structure of the human receptor, which includes the cysteine-rich linker region. The C-terminal kinase domain shares 45% identity with that of the human insulin RTK. Modeling was performed on the kinase domain, since both unphosphorylated and phosphorylated forms of the human protein have known structures: PDB entries 1irk and 1ir3, respectively (Hubbard, 1997).

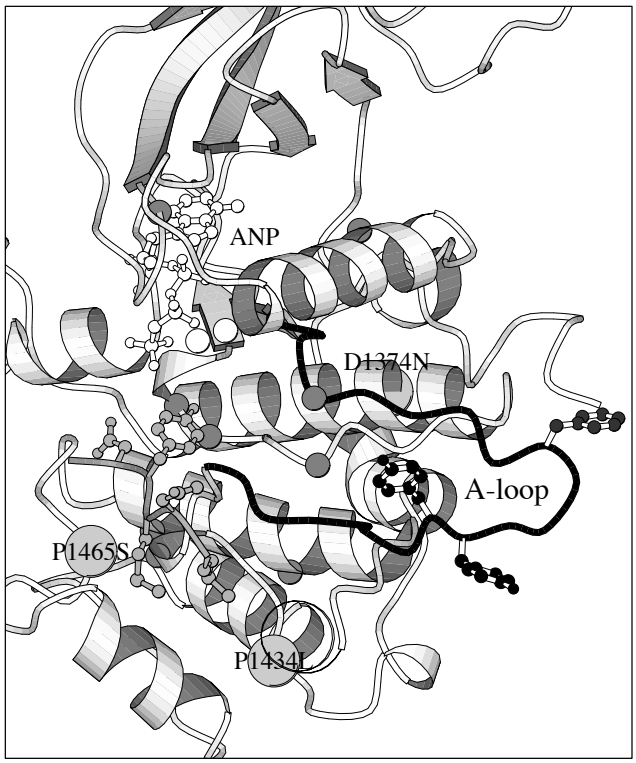

Figure 3. Model of DAF-2 showing ANP (a non-hydrolyzable ATP analog), two Mg^{2+} ions (white), and a peptide substrate (grey) bound at the active site. The activation loop (black) is shown with the tyrosines that become phosphorylated. Three daf-2 mutations are labeled, and several other mutations that have been found in human diabetic patents are shown as dark grey spheres.

The human tyrosine kinase domain has a nucleotide binding loop, catalytic loop, and activation loop (A-loop). The protein is activated in gradations, as successive tyrosines become phosphorylated (Hubbard, 1997). This leads to a 20-30 Å movement of the A-loop which stabilizes it and allows access to the active site by ATP and protein substrates (Figure 2). Full activation occurs once Y1163 (Y1419 in daf-2) is phosphorylated. The analogous tyrosines in daf-2 are Y1414, Y1418, and Y1419, located in the activation loop (residues 1405-1426). Y1414 and Y1418 may serve as docking sites for downstream signaling proteins (such as AGE-1, DAF-16, and DAF-18). AGE-1

shows homology to the catalytic subunit of mammalian phosphatidylinositol 3-OH kinase and is also related to longevity in the worm (as is apparent from its name).

Observed mutations that decrease DAF-2 signaling are shown in Figure 3 and include D1374N, P1434L, and P1465S. The proline mutations will disrupt or increase the mobility of the turns that they help form, which may affect local structure. We speculate that mutation of these conserved prolines has a destabilizing effect on the protein, rather than directly affecting ligand binding. The sidechain of D1374 is in the middle of a hydrophobic pocket and both delta oxygens hydrogen bond to the backbone amide of F1533. Replacement of one of these oxygens with nitrogen is apparently destabilizing enough to the overall structure to give an observable phenotype, even though D1374 is not strictly conserved among tyrosine kinases.

Interestingly, the P1434L mutation corresponds to a substitution in human insulin RTK observed in a diabetic insulin-resistant patient (Kimura *et al*., 1997). Many more mutations have been characterized in the human protein from non-insulin dependent diabetes patients (and are mapped onto the daf-2 model in Figure 3, shown as dark grey spheres). To show that these mutations are directly responsible for diabetes, it would at least be necessary to show that the mutation greatly impairs the functional activity of the protein in response to its substrate.

VAB-1

VAB-1 (Variable ABnormal) belongs to the Ephrin receptor tyrosine kinase (RTK) family (class 8). In *C. elegans*, the protein is involved in axon guidance and development of the nervous system. It is orthologous to mouse Nuk and human EphA (32% identity), which are involved in mammalian morphogenesis (Popovici *et al.*, 1999).

VAB-1 is a good example of the information that can easily be overlooked in BLAST results. For example, the BLAST synopsis in the WormPD only lists the Eph tyrosine kinase domain (30% of the protein), but most or all the major domains of the protein can be identified by PSI-BLAST (73% coverage).

A 200 residue segment in the N-terminus shows homology to the ligand-binding domain of the Eph receptor tyrosine kinase (24% identity to 1nukA (Himanen *et al.*, 1998) with an E-value of 1×10^{-75}). Another segment of over 200 residues is a fibronectin type III domain (21% identity to 1fnf with an E-value of 2×10^{-14}). Residues 650-1000 were modeled based on the human tyrosine kinase C-SRC (35% identity to 2src with an E-value of 2×10^{-68}). The last domain on the C-terminal end of the protein is the short Eph receptor SAM domain (20% identity to 1b0xA with an E-value of 1×10^{-9}). Domain assignments were confirmed by a biologist studying vab-1 and vab-2 (personal communication, Ian Chin-Sang).

The ligand-binding domain is shown as an example in Figure 4, modeled using the crystal structure of mouse receptor tyrosine kinase (RTK) EphB2 (Himanen *et al.*, 1998). The ligand-binding region is near the highly variable specificity loop, which packs against the concave surface of the β-sandwich scaffold. Mutations (E62K, T63I, and E195K) near this region of the protein have been shown to interfere with binding (George *et al*., 1998).

Models of the wildtype and mutant proteins were analyzed using GRASP (Nicholls *et al.*, 1991). The rendered electrostatic contours (data not shown) indicate that the environment experienced by a potential ligand is dramatically affected by the E62K mutation; the negative electric field at the top of the protein is disrupted by the solvent-exposed positive charge of the lysine sidechain. This finding explains the classification of this mutant as having a "strong" effect (George *et al*., 1998).

The T63I mutation has an "intermediate" effect, which can be explained by the more electrostatically conservative mutation of Thr63 to isoleucine. Finally, the E195K mutation introduces a larger, oppositely charged sidechain.

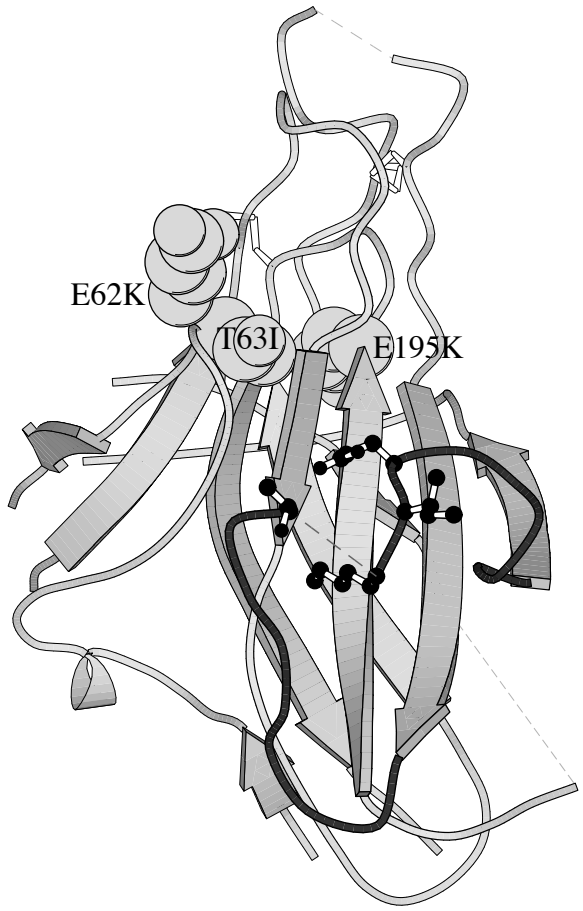

Figure 4. Model of the ligand-binding domain of *C. elegans* VAB-1. The ephrin class specificity loop is indicated in black, where binding occurs. Four residues (Tyr-Lys-Ile-Glu), unique to this subclass, are shown, modeled with SCWRL. The two disulfide bonds (top) are preserved in the model. Structure figures were made with MolScript (Kraulis, 1991).

MEL-32

MEL-32 is a serine hydroxymethyltransferase (SHMT) (Vatcher *et al.*, 1998), which converts serine to glycine. Mutations with an observable phenotype are lethal to developing embryos (MEL = Maternal Effect Lethal). The enzyme is highly conserved, with 61% identity to both human and plant SHMT's. Orthologs in yeast include Shm1p and Shm2p, although deletion of Shm2p in yeast is not lethal. Human tumors sometimes have elevated expression levels of SHMT, which is why it has been proposed as a chemotherapy target (Renwick *et al.*, 1998; Matthews *et al.*, 1998).

This protein provided a good comparison of the information content of various models based on two different template structures, an ornithine decarboxylase (1ord) with 12% sequence identity to MEL-32, and a newly solved human SHMT structure (1bj4) with 60% identity. The former is a gross "low resolution" model, whereas the latter is likely to be fairly accurate. The active site geometry of the two models is similar (see Figure 5), as are most regions of the overall fold. However, the model based on the ornithine decarboxylase obviously deviates from the correct MEL-32 fold in a number of regions.

MEL-32 is a PLP-dependent (pyridoxal 5'-phosphate) enzyme, so mutations that disrupt the active site and/or PLP binding should show a distinct phenotype, since the enzyme is essential for viability. A number of mutations have been characterized (Vatcher *et al.*, 1998), and many of them are localized in and around the active site (see Figure 6). For example, the R84Q mutation near the opening to the active site will disrupt positive charge that is essential for activity. G406E introduces a negatively charged sidechain into the active site while L146F sterically blocks access to the binding site. G149E may

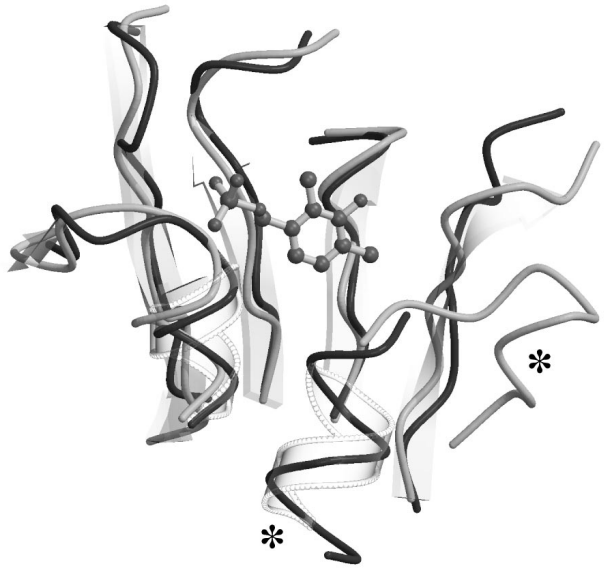

Figure 5. Active site superposition of the models of MEL-32 based on ornithine decarboxylase (dark) and hSHMT (light). Most of the backbone geometry is preserved, except in the area indicated by asterisks. The pyridoxal-5'-phosphate is shown in the active site.

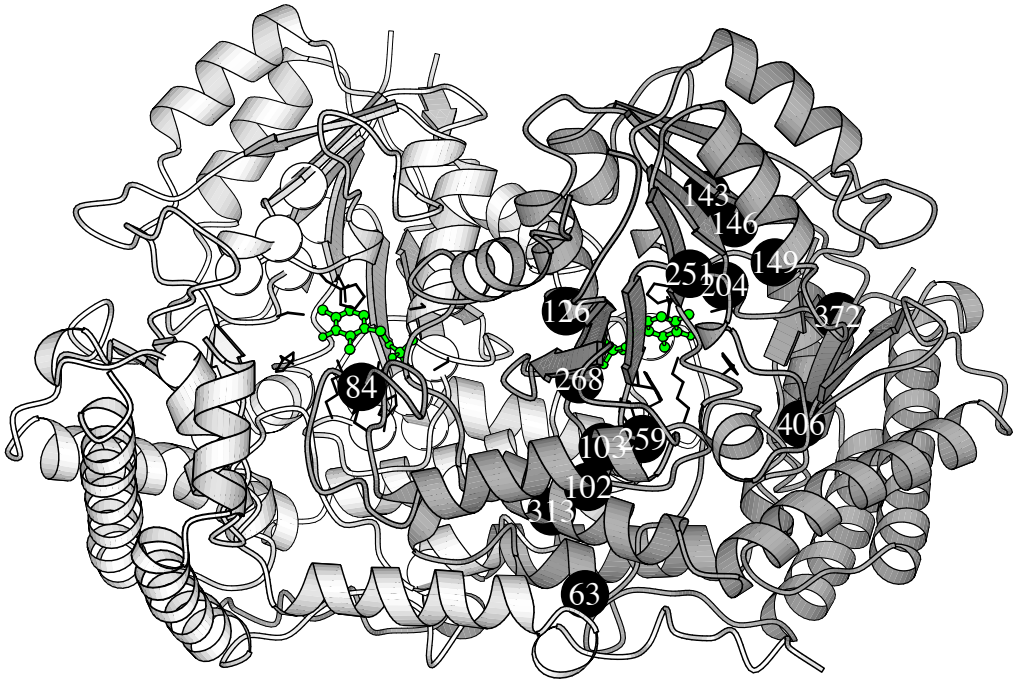

Figure 6. Model of the MEL-32 dimer, with PLP (ball-and-stick representation) in the active sites and mutations shown as large spheres. For clarity, mutations are black in one monomer and white in the other. Important active site residues are shown as wireframes.

distort a segment of the chain forming the active site since the backbone ϕ,ψ angles (110°,5°) of the conserved glycine are disallowed for glutamic acid. H259 is an important active site residue; mutation to Tyr will disrupt PLP binding.

The functionally active enzyme forms a dimer or tetramer (dimer of dimers), and a number of the other mutations affect crucial interactions between the two monomers. R102 makes a salt bridge to D33 and E35 on the adjacent monomer. The R102K substitution, though conservative, is apparently enough to weaken this salt bridge. Furthermore, the substitution of A63 by valine will force movement of V32 on the opposite chain, which then displaces D33 and likely alters the geometry necessary for salt bridge formation.

In some cases, compensatory mutations in different alleles (and consequently different monomers) allow a recovery of enzyme function, as indicated by surviving larvae. One of these cases of heterozygous alleles is t1597/t1576 (R84Q/G372R, personal communication with Greg Vatcher and Heinke Schnabel). Although many of these heterozygous mutations are difficult to explain without invoking an effect on tetramer formation, this example suggests that partial loss of charge near the active site of one monomer (R84Q) is rescued by introduction of a positive charge on the other monomer (G372R), but located in close proximity to the same active site.

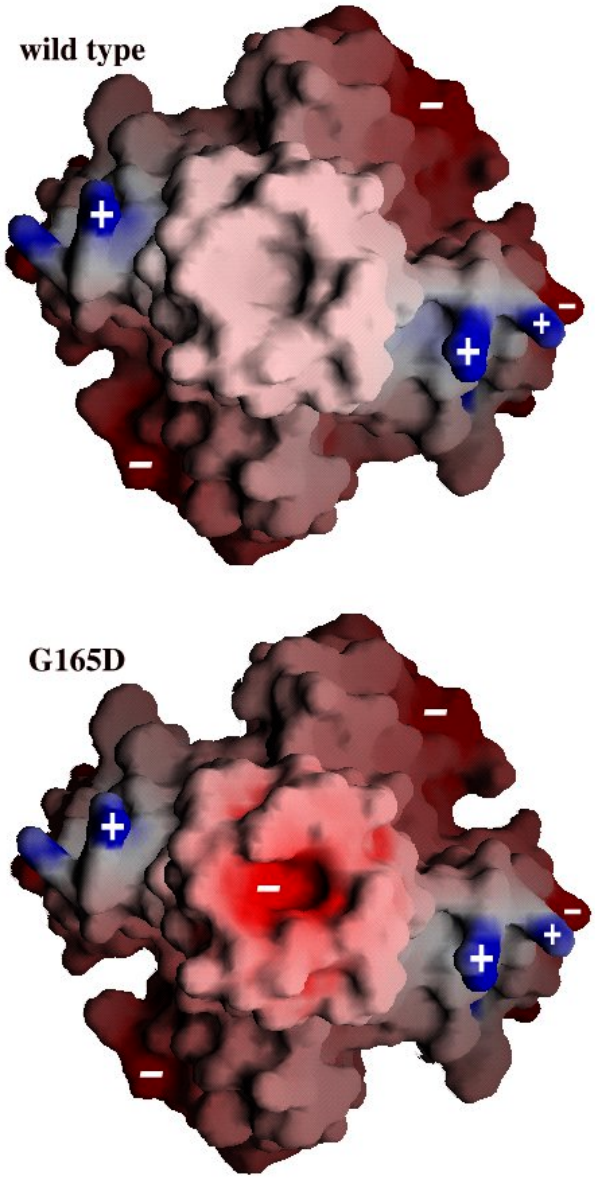

Figure 7. TWK-18 potassium channel model showing the effect of the G165D mutation on the electrostatic potential at one mouth of the channel.

TWK-18

TWK-18 belongs to the TWIK 4 transmembrane (TM) potassium channel family. There are many 4 TM potassium channel proteins in the C. elegans genome; they are homologous to potassium channel proteins in yeast, Drosophila, mouse and human, with 22-28% sequence identity.

The PDB template structure (1bl8) is a potassium channel from *Streptomyces lividans* (Doyle *et al.*, 1998), a tetramer of 4 distinct chains. The C. elegans protein, by contrast, is a dimer of 2 chains, each chain having two transmembrane segments. Potassium transport is mediated by backbone carbonyl oxygens lining the inside of the channel. The geometry of this "selectivity filter" confers the high degree of specificity that enables the channel to select for K^+ ions but not smaller Na^+ ions. A water-filled cavity inside the pore and helix dipoles directed toward the center of the pore provide the electrostatic environment necessary for the cation to surmount the energy barrier of crossing a membrane bilayer. A large patch of negatively charged sidechains on the extracellular surface of the protein provides an attractive force to the potassium cations.

One mutation at the mouth of the intracellular surface of the channel, G165D, has been characterized as a gain-of-function mutation (Maya Kunkel, personal communication). In an attempt to understand how this substitution might enhance conduction through the channel, the glycine was mutated to an aspartic acid in our twk-18 model using the SCWRL program. Analysis using GRASP (Figure 7) clearly indicates the effect of this negatively charged sidechain. We propose that this additional negative charge at the exit of the channel provides an additional electrostatic tug on the ion as it crosses from the water-filled cavity in the middle of the membrane to the intracellular exit port.

Alzheimer precursor protein β-secretase

Alzheimer's disease is characterized by a build up of plaque in the brain consisting primarily of a 42-46 amino acid peptide cleaved from the Alzheimer precursor protein (APP), a membrane-bound protein of unknown function.

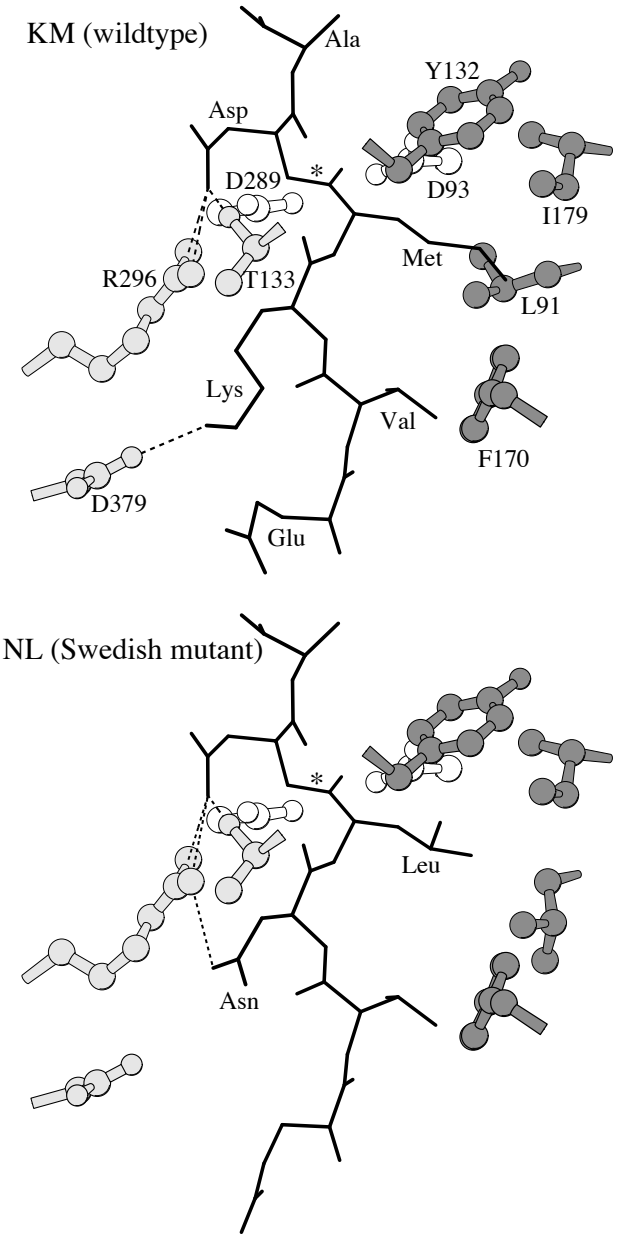

Figure 8. Models of β-secretase with APP-derived substrates. The top panel shows the wildtype substrate with sequence EVKMDA. According to standard protease nomenclature, this corresponds to P4-P3-P2-P1-P1´-P2´. The cleavage site, between P1 and P1', is indicated by an asterisk (*). In the lower panel, the Swedish mutant (NL) is associated with early-onset Alzheimer's disease.

The predominant form of the Aβ peptide represents amino acid 672-713 of APP, cleaved from the parent protein by two enzymes referred to as β and γ secretase at the N and C terminus of Aβ respectively. Recently, the gene and protein responsible for the β-secretase activity was identified independently by four research groups (Hussain *et al.*, 1999; Yan *et al.*, 1999; Vassar *et al.*, 1999; Sinha *et al.*, 1999). The protein variously called BACE, ASP2, and β-secretase is a 501 amino acid protein, with a single transmembrane domain. The external domain is homologous to aspartyl proteases such as pepsin, cathepsin, and renin. β-secretase is unusual for aspartyl proteases in that it cuts at sites in APP that are negatively charged. The main proteolytic site is between M671 and D672. There is also weak cleavage between Y681 and E682.

While no polymorphisms of BACE have been reported to date, variations in the APP sequence near the cleavage site do predispose patients to early development of Alzheimer's disease. Mutations of K670-M671 to N670-L671 just prior to the cleavage site have been associated with early onset Alzheimer's disease in Swedish families (Lannfelt *et al.*, 1994).

We undertook model building of the aspartyl protease domain of BACE to understand the specificity of the enzyme for APP and to explain the differential cleavage rates of APP polymorphisms. We used a crystal structure of pepsin (PDB entry 1PSO) (Fujinaga *et al*, 1995) to model the enzyme and a crystal structure of rhizopuspepsin with a reduced bond peptide inhibitor to model the substrate (Suguna *et al.*, 1987). This is the only aspartyl protease structure with an inhibitor that does not contain additional heavy atoms in the inhibitor backbone. This facilitates the modeling of a peptide substrate into the active site, without the need for rebuilding the backbone of the substrate. The sequence identity between BACE and pepsin was 24%. All insertions and deletions were far from the active site, and were not modeled. The conformations of sidechains from the enzyme and the substrate were modeled simultaneously with the SCWRL program. Predicted structures for the most common allele of the substrate (KM) and the Swedish mutation allele (NL) with active site residues of the enzyme are shown in Figure 8.

The specificity of BACE for negatively charged residues at P1´ was immediately obvious from the model. The buried R296 of BACE is in a position to form a salt-bridge with the partially buried Asp (or Glu) of the substrate. Met at the P1 position is surrounded by hydrophobic residues (L91, I179, Y132) as is valine at position P3. The lysine at P2 is able to form a salt-bridge with D379. Taken together, it is clear that APP is a good substrate for BACE. We also modeled the Swedish mutant, KM→NL at P2,P1 positions. It is clear the Asn residue can also hydrogen bond to R296 and that the Leu fits nicely into the hydrophobic pocket formed by L91, I179, and Y132. It is not immediately obvious why the Swedish mutant is a better substrate than the wildtype, but the specificity determining sidechains of the substrate fit nicely into the active site.

Cystathionine β-synthase

Elimination of the methionine metabolite homocysteine from the blood is accomplished in large part by the enzyme cystathionine β-synthase (CBS). High levels of homocysteine are strongly linked to heart disease (Refsum *et al.*, 1998), and patients with homocystinuria have been

found to have mutations in the gene that codes for CBS (Kraus *et al*., 1999). Most such patients are responsive to pyridoxal phosphate (vitamin B6), which is a coenzyme covalently linked to CBS. We modeled CBS in collaboration with Dr. Warren Kruger since it is very well characterized in terms of its polymorphisms in human populations, and represents a good test case for the interpretation of genetic variation in humans through comparative modeling.

Our first attempt at fold assignment for CBS produced a hit in the β chain of tryptophan synthase with a sequence identity of 18%. While the homology is low, both enzymes utilize pyridoxal phosphate, and catalyze very similar reactions. In the case of tryptophan synthase, the β chain replaces the hydroxyl of serine with indole to synthesize tryptophan. In the case of CBS, the enzyme replaces the serine hydroxyl with homocysteine to produce cystathionine. Cystathionine is subsequently converted into cysteine by cystathionine γ-lyase. Recently crystal structures of two enzymes more closely related to CBS were deposited in the PDB: threonine deaminase and O-acetylserine sulfhydrylase. The latter structure has a 38% sequence identity to CBS, and provides the best template for comparative modeling of CBS.

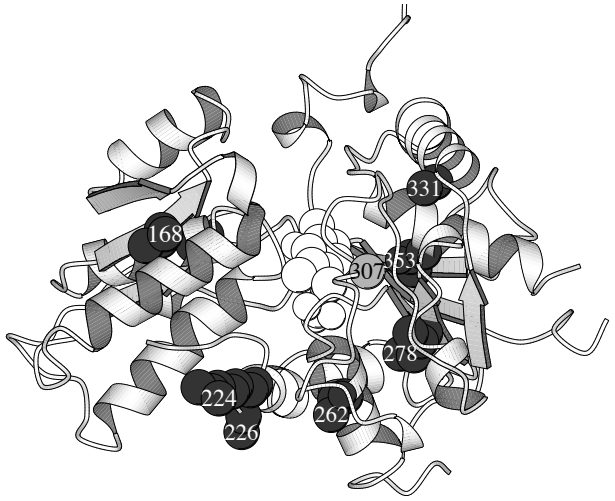

Figure 9. Model of cystathionine β-synthase. The bound pyridoxal phosphate is shown in white. Mutations are in black; the location of the G307S mutation is highlighted in grey.

Human CBS is a 501 amino acid protein. The first 75 amino acids form a proline rich region, followed by the enzyme domain homologous to tryptophan synthase, threonine deaminase, and O-acetylserine sulfhydrylase, all of which share quite similar folds. The enzyme domain of CBS is followed by a 155 amino acid region, that contains a 53 amino acid motif found in many proteins, including inosine monophosphate dehydrogenase, chloride channels, and 5´AMP-activated protein kinase γ subunit (Bateman, 1997). This domain is referred to as a CBS domain, in reference to its presence in CBS (residues 415-468 in human CBS). In inosine 5'-monophosphate dehydrogenase (IMPDH), two such domains are inserted in tandem between the second helix and third strand of the central α-β barrel of the enzyme. The function of the C-terminal region of CBS (residues 396-551) is not absolutely clear, but appears to act as a regulator of enzyme function: binding of S-adenosyl methionine to this domain activates the protein, and elimination of the domain produces an enzyme that is constitutively active. The most straightforward model is to hypothesize that the C-terminal domain provides gated access to the active site.

We modeled the central enzyme domain based on the crystal structure of O-acetylserine sulfhydrylase (Burkhard *et al*., 1998) (PDB entry 1oas). A large number of mutations have been observed in human patients with homocystinuria. These include A114V, V168M, R224H, A226T, T262M, I278T, G307S, V320A, A331V, and T353M which have been studied *in vitro* in yeast (Shan *et al*., 1998). In all cases except G307S, patients were responsive to vitamin B6 therapy. And in all cases except G307S, artificial constructs that eliminated the C-terminal domain created constitutively active proteins. These mutations have been mapped onto the model of CBS and shown in Figure 9. The model supports the hypothesis that capping of the active site by the C-terminal domain acts as a regulator of activity. The mutations found in B6-responsive patients are scattered about a single face of the protein, which forms the active site of the enzyme domain. It is likely that these mutations interfere with the conformational change that removes the C-terminal cap, perhaps increasing the binding affinity of the cap for the active site face of the protein. The alteration in the single B6-unresponsive allele, G307S, is directly in the active site of the enzyme. Building this amino acid into the model brings the mutant serine hydroxyl close to the pyridoxal phosphate. This is likely to interfere directly with pyridoxal phosphate binding and catalysis.

Conclusion

The availability of complete genomes has tempted computational biologists into creating databases of fold assignments and comparative models (Sanchez *et al*., 2000; Guex *et al*., 1997). Genome fold assignments are helpful in understanding ancient conserved regions, gene duplications, etc. (Brenner *et al*., 1995), but without interpretation of available experimental data, these models and assignments have little effect on our understanding of the biological function of particular genes and proteins. Molecular biologists are generally unskilled in looking at structure models and interpreting sequence changes in terms of electrostatic effects and changes in stability and dynamics. We feel that it is incumbent on the modeling community to take an active approach in choosing systems for study and in pursuing collaborations with experimental biologists. In this paper we have described six such examples.

After assessing the alignment accuracy of PSI-BLAST, we generated fold assignments for the complete *C. elegans*

genome using PSI-BLAST, as well as for all currently available human protein sequences in GenBank. We proceeded to make models of proteins in both genomes in consultation with biologists actively studying these systems. In most cases, the structural basis of mutations that change phenotype can be easily postulated in terms of the models. Quite frequently the most deleterious mutations lie in or around an active site, replacing conserved residues. Mutations to charged or polar amino acids in the hydrophobic core can also be explained in terms of protein stability. Mutations of conserved prolines, glycines, and disulfide-bonded cysteines are also likely to have a large effect on local structure and dynamics. In other cases, however, the mutations can not be easily interpreted. This may be because of missing data, such as the location of multimer interfaces, or because of quite subtle long-range effects on protein structure and dynamics. To understand these situations will require further experimental and computational work, including mutagenesis, structure determination, and molecular dynamics simulations (Zhou *et al*., 1999).

Notes

Several tools and supplementary information are available from our web site.

- The *S2C database* correlates the residue numbering in PDB SEQRES and ATOM coordinate records and flags errors in these records:
  http://www.fccc.edu/research/labs/dunbrack/s2c/
- *SCWRL* and the backbone-dependent rotamer library:
  http://www.fccc.edu/research/labs/dunbrack/scwrl/
- *Genomic fold assignments:*
  http://www.fccc.edu/research/labs/dunbrack/genomes/

Acknowledgments.

This work was funded in part by grant CA06927 from the National Institutes of Health, a grant from the American Cancer Society, and an appropriation from the Commonwealth of Pennsylvania. J.M.S. is supported by NIH Post-doctoral Training Grant CA09035 awarded to Fox Chase Cancer Center from the National Cancer Institute. We thank Jonathan Arthur for his work on the β-secretase project.

References

Altschul, S. F. and Koonin, E. V. 1998. Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends in Biochem. Sci*. 23:444-447.

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

Arthur, J. W. and Dunbrack, R. L., Jr. 2000. Correlating residue numbering in the Protein Databank. *submitted*.

Bateman, A. 1997. The structure of a domain common to archaebacteria and the homocystinuria disease protein. *Trends Biochem Sci* 22:12-3.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res*. 28:235-242.

Bower, M., Cohen, F. E. and Dunbrack, R. L., Jr. (1997). SCWRL: A program for building sidechains onto protein backbones. www.cmpharm.ucsf.edu/~bower/-scwrl.html, University of California San Francisco.

Brenner, S. E., Chothia, C. and Hubbard, T. J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* 95:6073-6078.

Brenner, S. E., Hubbard, T., Murzin, A. and Chothia, C. 1995. Gene duplications in H. influenzae. *Nature* 378:140.

Brenner, S. E., Koehl, P. and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 28:254-256.

Burkhard, P., Rao, G. S., Hohenester, E., Schnackerz, K. D., Cook, P. F. and Jansonius, J. N. 1998. Three-dimensional structure of O-acetylserine sulfhydrylase from *Salmonella typhimurium*. *J. Mol. Biol*. 283:121-133.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaastgerland, T., Lin, D., Sali, A., Studier, F. W. and Swaminathan, S. 1999. Structural genomics: beyond the Human Genome Project. *Nature Genetics* 23:151-157.

Costanzo, M. C., *et al*. 2000. The Yeast Proteome Database (YPD) and Caenorhabditis elegans Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res*. 28:73-76.

Domingues, F. S., Lackner, P., Andreeva, A. and Sippl, M. J. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol*. 297:1003-1013.

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. and MacKinnon, R. 1998. The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* 280:69-78.

Dunbrack, R. L., Jr. 1999. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins Suppl* 3:81-87.

Dunbrack, R. L., Jr. and Cohen, F. E. 1997. Bayesian statistical analysis of protein sidechain rotamer preferences. *Prot. Science* 6:1661-1681.

Fujinaga, M., Chernaia, M. M., Tarasova, N. I., Mosimann, S. C. and James, M. N. 1995. Crystal structure of human pepsin and its complex with pepstatin. *Protein Sci* 4:960-72.

George, S. E., Simokat, K., Hardin, J. and Chisholm, A. D. 1998. The VAB-1 Eph receptor tyrosine kinase functions in neural and epithelial morphogenesis in *C. elegans*. *Cell* 92:633-643.

Gerstein, M. 1998. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* 33:518-34.

Guex, N. and Peitsch, M. C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714-23.

Hamosh, A., Scott, A. F., Amberger, J., Valle, D. and McKusick, V. A. 2000. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* 15:57-61.

Himanen, J.-P., Henkemeyer, M. and Nikolov, D. B. 1998. Crystal structure of the ligand-binding domain of the receptor tyrosine kinase EphB2. *Nature* 396:486-491.

Holm, L. and Sander, C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* 26:316-9.

Hubbard, S. R. 1997. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* 16:5572.

Hussain, I., *et al.* 1999. Identification of a novel aspartic protease (Asp2) as β-secretase. *Mol. Cell. Neurosci.* 14:419-427.

Karplus, K., Barrett, C. and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14:846-856.

Kimura, K. D., Tissenbaum, H. A., Liu, Y. and Ruvkun, G. 1997. daf-2, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science* 277:942-946.

Koehl, P. and Levitt, M. 1999. A brighter future for protein structure prediction. *Nature Struct. Biol.* 6:108-111.

Kraulis, P. J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24:946-950.

Kraus, J. P., *et al.* 1999. Cystathionine beta-synthase mutations in homocystinuria. *Hum Mutat* 13:362-75.

Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeysinghe, S., Thomas, N. and Cooper, D. N. 2000. Human gene mutation database—a biomedical information and research resource. *Hum. Mutat.* 15:45-51.

Lannfelt, L., Bogdanovic, N., Appelgren, H., Axelman, K., Lilius, L., Hansson, G., Schenk, D., Hardy, J. and Winblad, B. 1994. Amyloid precursor protein mutation causes Alzheimer's disease in a Swedish family. *Neurosci Lett* 168:254-6.

Matthews, R. G., Drummond, J. T. and Webb, H. K. 1998. Cobalamin-dependent methionine synthase and serine hydroxymethyltransferase: targets for chemotherapeutic intervention? *Adv. Enzyme Regul.* 38:377-392.

Nicholls, A., Sharp, K. and Honig, B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11:281-296.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284:1201-1210.

Popovici, C., Roubin, R., Coulier, F., Pontarotti, P. and Birnbaum, D. 1999. The family of Caenorhabditis elegans tyrosine kinase receptors: similarities and differences with mammalian receptors. *Genome Res.* 9:1026-1039.

Refsum, H., Ueland, P. M., Nygard, O. and Vollset, S. E. 1998. Homocysteine and cardiovascular disease. *Annu Rev Med* 49:31-62.

Renwick, S. B., Snell, K. and Baumann, U. 1998. The crystal structure of human cytosolic serine hydroxymethyltransferase: a target for cancer chemotherapy. *Structure* 6:1105-1116.

Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P. I., Wittenstein, E. and Sali, A. 2000. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* 28:250-253.

Sauder, J. M., Arthur, J. W. and Dunbrack, R. L., Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40:6-22.

Sayle, R. A. and Milner-White, E. J. 1995. RASMOL: biomolecular graphics for all. *Trends in Biochem. Sci.* 20:374.

Shan, X. and Kruger, W. D. 1998. Correction of disease-causing CBS mutations in yeast. *Nat Genet* 19:91-3.

Shindyalov, I. N. and Bourne, P. E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Prot. Eng.* 11:739-47.

Sinha, S., *et al.* 1999. Purification and cloning of amyloid precursor protein beta-secretase from human brain. *Nature* 402:537-540.

Suguna, K., Padlan, E. A., Smith, C. W., Carlson, W. D. and Davies, D. R. 1987. Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. *Proc Natl Acad Sci U S A* 84:7009-13.

Vassar, R., *et al.* 1999. β-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science* 286:735-741.

Vatcher, G. P., Thacker, C. M., Kaletta, T., Schnabel, H., Schnabel, R. and Baillie, D. L. 1998. Serine hydroxymethyltransferase is maternally essential in *Caenorhabditis elegans*. *J. Biol. Chem.* 273:6066-6073.

Wolf, Y. I., Brenner, S. E., Bash, P. A. and Koonin, E. V. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* 9:17-26.

Yan, R., *et al.* 1999. Membrane-anchored aspartyl protease with Alzheimer's disease beta-secretase activity. *Nature* 402:533-7.

Zhou, Y., Vitkup, D. and Karplus, M. 1999. Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. *J. Mol. Biol.* 285:1371-5.