

BIOCHEM218: Computational Molecular Biology
Professor Douglas Brutlag
Charles Kou
charlesk@stanford.edu
March 15, 2005

Analysis of Protein Structure Prediction by Homology Modeling

Abstract

The method of protein structure prediction is critically reviewed with emphasis on homology modeling. The current state of the art techniques and limitations are analyzed and possible improvements are suggested.

Introduction

Protein structure prediction is one of the most important problems of computational molecular biology. The accurate prediction of the protein three-dimensional structure (tertiary structure) from the amino acid sequence (primary structure) could facilitate rational drug design. In rational drug design, the ability to predict the tertiary structure of the protein from the sequence could facilitate researchers to design drugs that can specifically target the key molecule to stop the functioning of the pathway in the diseased state or enhance the functioning of the pathway inhibited by the diseased state. This requires a very high accuracy and high-resolution model to be useful. On the other hand, lower resolution model could still give insight into the function of the unknown sequence, help design molecular biology experiments, and guide cloning and purification design.

Human Genome Project, founded in 1990 by NIH and Department of Energy, is a high-throughput sequencing which produced plethora of information that are made available on the Internet database for public use. The high volume sequencing data created by the advancement in computational biology created a lag between the availability of sequence data and the determination of the three-dimensional structure of the corresponding sequences. Experimentally determining the three-dimensional structure of the protein sequence through x-ray crystallography or NMR spectroscopy is expensive and time-consuming. Protein structure prediction can serve the need of the scientific community by providing an efficient alternative to determining protein structure of the high-throughput sequence data produced by the Genome Project.

CASP

Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a contest for protein structure prediction, which began in 1994 as CASP1 and subsequently held every two years as CASP2 (1996), CASP3 (1998), CASP4 (2000), CASP5 (2002), and CASP6 (2004). The latest CASP6 took place in December 4-8, 2004 in Gaeta, Italy. The latest published result of CASP5 appeared in Proteins: Structure, Function, and Genetics Volume 53, Issue S6. The contest aims to compare various methods for the advancement of the field of protein structure prediction.

The competition is framed so that the three main subfields of protein structure prediction can be advanced. Sequences of the unknowns, which are categorized based on

the similarity to the already existent three-dimensional models are made available to the entrants. The degree of similarity determines the category and the algorithm used to predict the model. When there is a high percentage of similarity to the already existent model, homology modeling is used. When there is a low similarity, fold recognition/threading is used. Lastly, when there is little correlation between the currently available model and the sequence, ab initio method is used.

Experimentalist determines the structures of the sequences by x-ray crystallography or NMR spectroscopy. The predictors use the sequences and apply their implementation of protein prediction algorithm in one of the various categories based on the type of the unknown sequence. Lastly, assessor will analyze the quality of predictions by comparing the experimentalist's experimental result and the predictor's theoretical model using criteria such as RMSD, overall identification, and topology, energy considerations such as contacts, H-bonds, similarity of the hydrophobic core, and the sequence alignment quality.

Challenges of Protein Structure Folding (Protein Structure Prediction via Ab initio)

Protein structure prediction is distinct from protein folding problem (ab initio), as folding problem is concerned with modeling and predicting the three-dimensional structure from primary structure using physical principles. On the other hand, protein structure prediction combines the use of statistical and experimental data to heuristically predict and refine the model. Only when protein structure prediction techniques such as homology modeling and fold recognition/threading fails, one resort to predicting the structure based on physical principle alone (ab initio / protein folding problem).

Modeling of protein structure folding is very difficult given the current state of computational power and the lack of complete theoretical framework. To model the structure, constituent atoms, bond length, bond angles, and constraints on dihedral angles must be considered. The size of the state space of specific three-dimensional conformation is large, because the bond between the neighboring amino acids can be bent and twisted in various ways. If one assumes the state space is searched in a sequential search, the theoretical calculation would take much longer than the actual time span of few milliseconds that atoms take to minimize the energy state. The fact that primary sequence alone does not fully specify the tertiary structure makes the problem more difficult. For example, chaperonins can induce proteins to fold in specific ways, and primary solvent (water or lipid), the concentration of salts, temperature and other environmental factors can affect the folding. Tertiary structure also involves covalent bonding through disulfide bridge between two cysteines. In addition, hydrogen bonding, Van der Waals interactions also participate in the formation of tertiary structure.

Increased computing power is needed to solve the protein-folding problem. Current approach to the problem is to develop a super computer (Blue Gene) or use a distributed computing (Folding@Home) platform. Blue Gene is, as of November 2004, ranked as the world's most powerful super computer and provides sustained performance of 70.72 Teraflops. Folding@Home, on the other hand, utilizes the vast unutilized computing resources available on the Internet. The program runs as a screensaver on users' computer and provides computing power and the results of distributed calculation are sent to the central server. Collaboration with Google is expected to provide wider user base than currently available. Both approaches will provide ways to quicken

computation, but theoretical breakthrough and improved algorithm is essential in solving the folding problem.

For improved algorithm, one could learn from the theoretical framework gained from the experimental data. Protein folding in nature seems to progress by first establishing secondary structure (alpha helices, beta sheets, coils and loops), and following with the tertiary structure production. Therefore, to simplify the folding problem, one approach is to first convert the primary structure to the secondary structure, then build the tertiary structure by examining the interaction among the secondary structures. One could also use the Ramachandran plot to exclude some states as impossible states due to the space filling nature of side chains. Finally, Gibbs Free Energy function ($\Delta H - T \cdot \Delta S$) can be used as a guiding function. This is a good function to use, because it tends to bring hydrophobic residues inward while bringing hydrophilic ones outwards, resulting in higher degrees of freedom for surrounding water molecules because the favorable interaction of ΔH outweighs the cost of ΔS . When protein folds, the atoms are constrained in a particular state, and therefore the entropy is decreased. The use of these constraints can reduce the state space of possible atomic coordinates and make the search problem more tractable. Rosetta method is an example of algorithm used to tackle the ab initio problem.

Currently, due to the difficulty of the problem and the lack of computing power, the use of ab initio is limited to modeling short sequences. The ability to predict the protein structure for larger protein requires better algorithm, improvement in the theoretical framework, and increase in computational power. This is an important subfield of protein structure prediction because it could improve the accuracy of homology modeling and fold recognition by providing additional information to the process of template creation and model refinement.

Homology Modeling

Homology modeling is based on the assumption that the sequence that is > 25-30% similar to already known structure is highly likely to share the similar tertiary structure. Therefore, already existent three-dimensional model in Protein Data Bank (PDB) is used as a template and is used to predict the tertiary structure of the given sequence. The initial step is to use the sequence comparison database to find homologues. The homologues are then used to identify the template. The template is aligned with the given sequence, and a new model is created by computationally mutating the structurally divergent regions (SDR) to amino acid sequence corresponding to the unknown sequence. The side chain conformations are added, then the model is refined and evaluated (Diagram 1).

The accuracy of modeling ultimately depends on the quality of alignment used to determine the template and the final model that is created. In addition, the percentage of similarity (structural conservation) between the template and the unknown sequence is the key factor. Lastly, the predictive capability of SDR region and the placement of side-chain also will affect the accuracy because template cannot be used to determine the placement of these regions. The models are usually sufficiently accurate because most biologically important regions are conserved and therefore similar to the structural template, when good template match is found.

The accuracy of the method is expected to increase as more protein structures become available. The increase in the knowledgebase allows increase in the probability of finding a better homologue; this will allow matching of a better template that can be used to create a model. In addition, increase in the sequence database will also allow better detection of homologous relationships through techniques such as multiple-sequence alignments, profiles, and Hidden Markov Models (HMM). The increase in the knowledge is also expected to allow researchers to develop new and better methodologies of inferring the homologues.

Homology Modeling: Sequence Alignment

Finding homologue in the database, such as Protein Data Bank, SCOP, DALI, GenBank, GeneCensus, MODBASE, PRESAGE, SWISSPROT+TrEMBL and CATH is the initial step of protein structure prediction. There are many techniques for doing this. BLAST is a pair wise comparison which can detect sequence similarities of >30%. Multiple alignments can be also used, such as HMM and Profile. Another approach involves the use of motif and the use of “signatures” to search for the alignment such as eMOTIF. Pfams, PRINTS and BLOCKS can also increase the chance of finding remote homologues that cannot be easily detected using pair-wise alignment such as BLAST.

Multiple alignments can be used to increase the probability of the match. PSI-BLAST first builds profile by searching the database using the unknown sequence, and by iteratively searching the database using the search result, it attempts to increase the accuracy of the search result. HMM on the other hand creates a Hidden Markov Model for the unknown sequence through multiple alignments and uses the HMM to search the database for additional matches. These multiple-alignment methods outperforms pair wise techniques for sequences with similarities that drops below 25%.

Finding good homology is crucial as subsequent steps of protein structure prediction depends on the template being used. When there are multiple candidates for templates selection, creating a phylogenic tree can help in selecting template from the subfamily that is most similar to the unknown sequence. The surrounding environment for the template should also be compared to that of the unknown sequence. Lastly, the quality of the template can also affect the decision process. When there is no match, homology-modeling method must be abandoned in favor of ab initio or threading method.

Homology Modeling: Unknown Sequence – Template Alignment

Once the template is selected, an optimal alignment of the template and the sequence must be made. Here, the identity of the unknown sequence and the template also plays a role as similarity of over 40% gives high accuracy of alignment. Algorithms such as CLUSTAL, BLOCK or FASTA are used in this stage. Often, multiple structures and templates are used to create increase the accuracy of the alignment. The use of multiple structures allows better prediction and reduce gaps in secondary structure elements, in buried regions. Sometimes visual inspection and human intervention is necessary to improve the accuracy of the alignment. The ability to fully automate the human-intervention step is one of the goals of the CAFASP. When it is difficult to determine the best alignment of the template and the sequence, the 3D model is generated and the model is evaluated rather than determining the alignment accuracy.

Homology Modeling: Generation of Model

Once an appropriate template is found, fairly accurate model can be constructed using homology modeling algorithms. There are three major algorithms classes, which are all similar in the accuracy of the modeling given the proper template. In other word, the accuracy of the modeling depends on the accuracy of the initial template input. The modeling algorithm should be fast, accurate, easy to automate, and allows incorporation of external data (such as secondary structure, and experimental data).

MODELLER is an example of algorithm that satisfies the spatial restraints. It utilizes the given unknown sequence and matched homology three-dimensional structure to predict the unknown protein structure. It first collects distance distributions between atoms in given known protein structure. Then it utilizes the collected distribution to compute the positions for equivalent atoms in alignment, and finally, the result is refined using energetic, such as restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom-atom interactions due to force field. MODELLER uses real-space optimization method where the initial model is built using the distance and dihedral angle restraints based on the template structure, which is subsequently optimized using the constraints. This is more efficient than the distance geometry approach, where all lower and upper bounds models are constructed based on distances and dihedral angles variance.

COMPOSER is an example of modeling by rigid bodies. This algorithm dissects the protein folds into core regions, variable loops and side chains. The coordinates of the carbon atoms of conserved regions are calculated by averaging the template structures. The main carbons are generated by using the template with highest similarity. Loops are generated and appended by searching the database to identify region that is similar to the environment of the template. The side chains are added based on the energetic and the template conformation. Lastly, the model is refined by minimizing the energetic.

SEGMOD is an example of algorithm that utilizes segment matching or coordinate reconstruction. In this algorithm, the carbon atoms are used as guiding positions and the database is searched to find matching segments that are then fit into the guiding position to generate the model.

Regardless of the algorithm chosen, ultimately, the accuracy of modeling is dependent on the sequence identity of the unknown sequence and the given template. This is understandable given the way algorithm functions by basing the new model's distance distribution using the template distribution. Of the three algorithmic approaches, modeling by satisfaction of spatial restraints seems to be the most promising of all because it allows constraints derived from experimental data to be incorporated into the algorithm.

Homology Modeling: Modeling Loop

Accurately modeling loops is necessary for determining the functional specificity of a protein. For example, the exposed loop that resulted from deviation in the unknown sequence from the template can contribute to active and binding sites, which can determine the binding specificity of antigens by immunoglobulin. Therefore, accuracy in loop modeling is favored.

Loop modeling can be construed as a subset of protein folding problem. When the residue is longer than 5 sequences long, the problem becomes difficult. The fold is

influenced by the core regions and also by the sequence of the loop. One can approach this problem by applying the same technique used in predicting the protein structure.

Ab initio method is essentially a search problem, which seeks the state that minimizes the energy function. Representation of the state, the energy function and the search algorithms can be varied for optimal result. Some examples of search algorithms used are Monte Carlo with simulated annealing, biased probability Monte Carlo search, and searching through discrete conformations by dynamic programming. Monte Carlo algorithm essentially randomly samples the search space and at the end of simulation, the ensemble of randomly chosen points gives information about the search space. Similarly, dynamic programming is the algorithm that is also used for Needleman / Wunsch sequence alignment technique. Each residue is represented by finite number of discrete states, and the local minima of energy function is sought through dynamic programming algorithm. Degrees of freedom of representation can be varied, such as Cartesian coordinates, or dihedral angles, which can be optimized in continuous or discrete spaces. Loop prediction algorithms can be applied to model the interaction of several loops and loops interactions with ligands.

Another approach to loop modeling is through the use of database search for similar configurations. The stems, which are the atoms that precede and follow the actual loop are searched and the output of the search are filtered according to geometric configuration and sequence similarity. The result is superposed and refined using energy function. This approach is limited by the length of segment, because as the length increases, the amount of search space increases and subsequently the probability of hit is reduced.

Fold Recognition / Protein Threading

There are more than 3000 different structural folds as reported by CATH database (Diagram 2). When homology-modeling algorithm fails to return a matching template, which typically occurs when there are less than 30% match between the given sequence and homology, the sequence is matched against the folds database to see if any of the sequence can be adopted as a template. The secondary structure of the sequence is predicted and that knowledge is used to match with the folds database that is promising. Then, the template is aligned with the unknown sequence, and the tertiary structure is modeled using the algorithms discussed above. When fold recognition fails, ab initio method is utilized to predict the tertiary structure from the unknown sequence alone.

Discussion

The accuracy of the final model is dependent on the quality of the template. Therefore, the presence of an appropriate template in the database is a necessity. To improve the probability of making a match during the initial sequence alignment phase, there should be a coordinated effort to experimentally determine the tertiary structure of sequences that has low homology, so that the protein database can have representative tertiary structures available.

In addition to improving breadth of availability in the protein database, sequence matching algorithms can be further refined so that a good match can be made. Current use of PSI-BLAST and HMM offers relatively good result because both methods utilizes multiple sequence alignment to increase the probability of finding a good homology

match. Multiple sequence alignment is dependent on quality of the initial multiple sequence inputs. Sometimes, protein family information can be used to provide good initial multiple sequence data. Often, human intuition and intervention helps with selection of sequences that should be included in the creation of the profile. Therefore, to fully automate protein structure prediction with high accuracy and good result requires an improved way to imitate the “human intuition” and knowledge, so that the sequence alignment step can be automated.

When a matching template is found, the unknown sequence and the template must be aligned. Again, this step is dependent on the selection of a good template. When the unknown sequence and the template is similar, they can be aligned with a high degree of accuracy. When the similarity is low, the alignment becomes problematic. Given the fact that there are more sequences than the experimentally determined models, there is a high probability that the “best” template is still not very similar to the unknown sequence. If experimentally determining the structure is not an option, one could attempt to use multiple templates, since accuracy of homology modeling is dependent on the degree of template and unknown sequence similarity. It is difficult to make a good alignment when the similarity is low between the template and the unknown sequence. Therefore, creating many three dimensional candidates and screening out the model with the best energetic may be an alternative way of approaching the problem. Lastly, advancement in the protein-folding field could potentially solve the problem of low similarity template. If ab initio method becomes computationally feasible and accurate, one could choose to model via ab initio when homology modeling fails to find a template with high degree of similarity.

Once the template and the sequence are aligned, modeling algorithm can create the tertiary structure. This step is highly dependent on the accuracy of the previous steps: the finding of good template match, and the successful alignment of the unknown sequence and the template. The template provides the initial framework of the model, and therefore the resulting model’s main skeleton does not deviate from the template very much. Therefore, the improvement in the modeling step depends on the accurate placement of the side chains and the accurate prediction of the loop placement. Essentially, once a good template is found, the general framework of the protein is formed. Therefore, the advancement in this field benefits from the improvement of ab initio technique. A better way to model the interaction of the side chain with the template-based framework combined with the database search technique for finding side chains with similar environmental configuration could improve the accuracy of the modeling step.

Protein structure prediction is composed of four steps: template selection, template-sequence alignment, modeling, and evaluation. These steps all benefit from human intervention, especially the template selection and alignment stage, as these two steps are crucial in creating an accurate model. Therefore, fully automatic protein structure prediction requires a way to imitate the “human intuition” of experts, which is crucial in creating a good protein model.

Another approach to improving the result of the model is to combine techniques from the three categories. A “Frankenstein’s Monster” from CASP5 is an example of this approach. As this approach shows, the distinction between the three subcategories is beginning to blur. Instead of using one template, multiple templates are used to pick and

choose the “good segments.” These segments are stitched together, modeled and evaluated. Then, the information from secondary structure prediction is used to improve the structure of the model. Lastly, the improved models are stitched together to generate another template. The missing region in the stitched together template is provided by the ab initio method. This final template is utilized to model and the energetic is minimized. This method requires expertise and human intervention. Therefore, more work is needed to codify the human intuition and to automate the process.

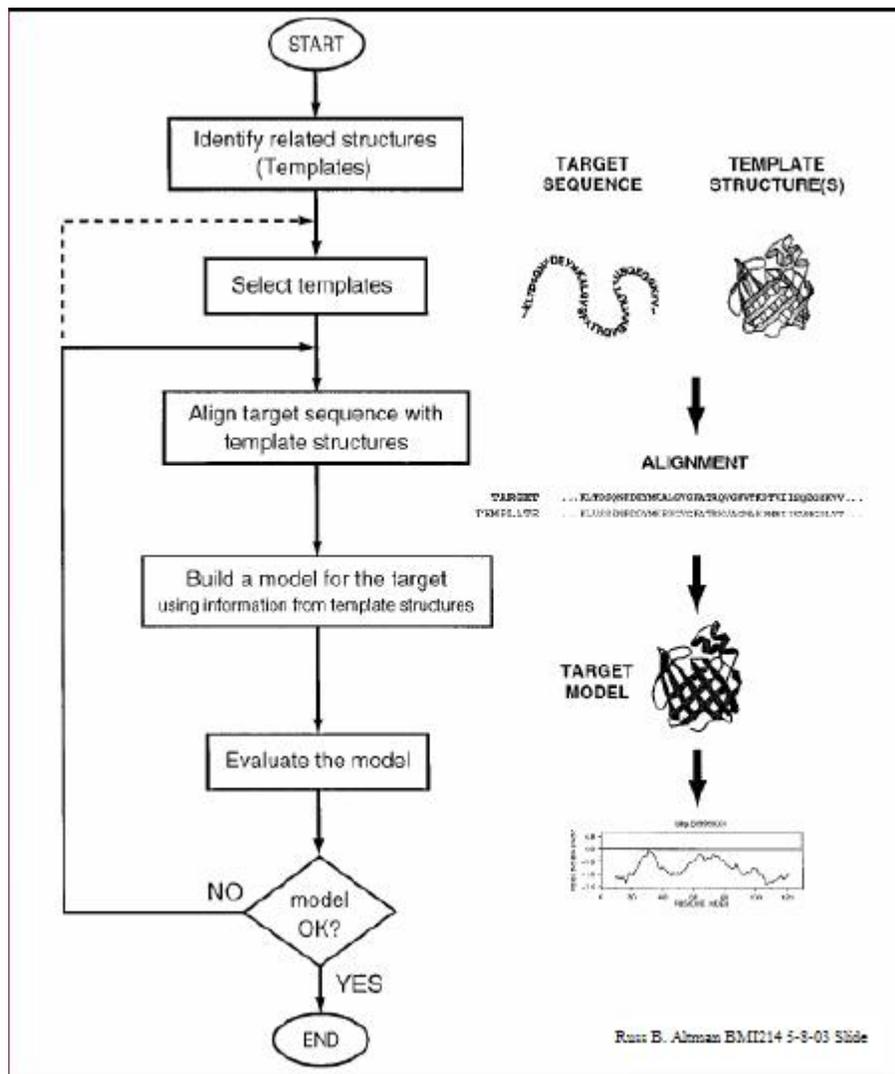


Diagram 1. Predicting the Model Using Homology.

CATH v2.5.1

| Version | 2.5.1 | | | | | | |
|---------------------------------------|------------|------|------|------|------|------|-------|
| Date | 28-01-2004 | | | | | | |
| C | A | T | H | S | N | O | D |
| Mainly Alpha | 5 | 227 | 428 | 948 | 1713 | 3946 | 10155 |
| Mainly Beta | 19 | 139 | 292 | 951 | 2344 | 5011 | 14259 |
| Alpha Beta | 12 | 368 | 648 | 2010 | 3631 | 8639 | 23025 |
| Few Secondary Structures | 1 | 86 | 91 | 114 | 225 | 378 | 952 |
| Multi-domain chains | 1 | 1053 | 1057 | 1071 | 2186 | 5801 | 12471 |
| Preliminary single domain assignments | 1 | 371 | 374 | 422 | 479 | 789 | 1663 |
| Multi-domain domains | 2 | 31 | 31 | 49 | 67 | 139 | 287 |
| CATH-35 Sequence families | 1 | 997 | 997 | 997 | 1108 | 2154 | 3431 |
| Fragments from multi-chain domains | 1 | 28 | 28 | 30 | 33 | 56 | 106 |

Diagram 2. Structural Folds at CATH
(<http://www.biochem.ucl.ac.uk/bsm/cath/releases.html>)

Works Cited

A. Fiser, R.K. Do, & A. Sali. Modeling of loops in protein structures, *Protein Science* 9. 1753-1773, 2000.

Altschul SF, Madden TL, Schaffer AA, Zhang J Zhang, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-402

Bajorath J, Stenkamp R, Aruffo A. 1994. Knowledge-based model building of proteins: Concepts and examples. *Protein Sci.* 2:1798 810

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235-42

Collura V, Higo J, Garnier J. 1993. Modeling of protein loops by simulated annealing. *Protein Sci.* 2:1502 10

Evans JS, Mathiowetz AM, Chan SI, Goddard WAIII . 1995. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology.

Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of

MCP603 from many randomly generated loop conformations. *Proteins* 1:342–62

Gribskov M. 1994. Profile analysis. *Meth. Mol. Biol.* 25:247–66

Henikoff S, Henikoff JG. 1994. Protein family classification based on searching a database of blocks. *Genomics* 19:97

Holm L, Sander C. 1996. Mapping the protein universe. *Science* 273:595–602

Holm L, Sander C. 1999. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*

Hubbard TJP, Ailey B, Brenner SE, Murzin AG, Chothia C. 1999. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* 27:254–56

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* 235:1501

Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226:507–33

M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291-325, 2000.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct.* 2000;29:291-325

Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin AC, Conte L .Lo, Thornton JM. 1999. The CATH database provides insights into protein structure/function relationship. *Nucleic Acids Res.* 27:275–79

Pearson WR. 1995. Comparison of methods for searching protein sequence databases. *Protein Sci.*

Rufino SD, Donate LE, Canard LHJ, Blundell TL. 1997. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modeling. *J. Mol. Biol.* 267:352–67

Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815

Sánchez R, Sali A. 1997. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins*

Sánchez R, Sali A. 1997. Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol.* 7:206–14

Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins. Part I. Three dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1:377–84

Vajda S, DeLisi C. 1990. Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers* 29:1755–72

Vlijmen HWT, Karplus M. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* 267:975–1001

Zhang ZT. 1997. Relations of the numbers of protein sequences, families and folds. *Protein Eng.* 10:757–61

http://en.wikipedia.org/wiki/Disulfide_bond

<http://en.wikipedia.org/wiki/Chaperonins>

http://en.wikipedia.org/wiki/Protein_structure_prediction

<http://en.wikipedia.org/wiki/Bioinformatics>

http://en.wikipedia.org/wiki/Protein_folding

http://en.wikipedia.org/wiki/Rational_drug_design

<http://en.wikipedia.org/wiki/CASP>

http://en.wikipedia.org/wiki/X-ray_crystallography

http://en.wikipedia.org/wiki/Nuclear_magnetic_resonance

http://en.wikipedia.org/wiki/Folding_at_Home

http://en.wikipedia.org/wiki/Blue_Gene

<http://folding.stanford.edu/>

<http://www.research.ibm.com/bluegene/>

http://www.wired.com/wired/archive/9.07/blue_pr.html

<http://www.smi.stanford.edu/projects/helix/bmi214/5-06-03clr.pdf>

<http://www.smi.stanford.edu/projects/helix/bmi214/5-8-03clr.pdf>

<http://predictioncenter.llnl.gov/>

<http://www.biochem.ucl.ac.uk/bsm/cath/releases.html>