

# Advanced Protein Modeling

HOMOLOG FINDING, SEQUENCE ALIGNMENT AND COMPARATIVE MODELING



Advanced Protein Modeling provides protein homolog searching, sequence alignment, and comparative protein modeling capabilities. The program recognizes distant homologs by sequence structure comparison and builds protein structures using information from one or more homologs. All steps in the modeling approach use information about how the substitution of amino acids are constrained by their local structural environments, which are defined in terms of secondary structure, solvent accessibility, and hydrogen bonding status.

## The Problem

The availability of complete genomes of many organisms has focused attention on the functions of the gene products, the proteins. This function is usually facilitated by the protein's accurate three-dimensional structure. Protein structures are determined experimentally by X-ray crystallography or NMR analysis. These methods are time consuming and frequently don't work for certain proteins. Since the three-dimensional fold within protein families is more conserved than their sequences, comparative protein modeling offers a solution to determine protein structures from sequence information *in silico*.

## Sequence-Structure Homology Recognition and Alignment<sup>1</sup>

The recognition of homology between protein sequences and known structures provides the foundation to understand biological behavior and is a prerequisite to the prediction of three-dimensional structures with comparative protein modeling.

Substitutions of amino acids in a protein structure are constrained by their local structural environment. Mainchain conformation or secondary structure, solvent accessibility and hydrogen bonding status have been shown to be useful structural features to describe local environments<sup>2</sup>.

The program uses 64 environment specific substitution tables which have been derived from structural alignments in the HOMSTRAD<sup>3</sup> database. A global-local algorithm is used to align a sequence-structure pair when they greatly differ in

length and a global algorithm is used in other cases. The respective alignment algorithm is selected automatically, which in itself improves alignments over other methods.

The gap penalty at each position of the structure is determined according to its solvent accessibility, its position relative to the secondary structure elements and the conservation of the secondary structure elements. The program is capable of aligning both multiple sequences and multiple structures to enrich the information about conservation and variation.

All of the above features have been demonstrated to improve both recognition and alignment accuracy of distant homologs.

## Model Structurally Conserved Regions<sup>4</sup>

The program identifies regions of homologous proteins of known structure that are likely to be structurally similar to regions of the target protein. It is a

### Selected Features

- Curated protein database with family structural profiles.
- Homolog detection using structural profiles and automatic alignment algorithm selection.
- Interactive sequence alignment editor links 1D sequences to 3D structures.
- Structurally conserved clusters expand the concept of structurally conserved regions.
- Rich graphical user interface supporting flexible workflows.
- Automatic project notes generation and capture.



Comparison of homology model of human factor Xa (Stuart-Prower factor, white) to the crystal structure (2BOK.pdb, orange), created using Advanced Protein Modeling. Sequence identities between the target and the five homologs used in modeling ranged between 40%-80%. The RMSD between the backbones of the model and the crystal structure is 1.5 Å.

fragment assembly approach which overcomes the problem that classical structurally conserved regions are defined as regions where all proteins of the same family show the same conformation for the main chain atoms, independent of their classification in a secondary structure element or loop region.

Advanced Protein Modeling adopts the concept of structurally conserved clusters, which use as much of the information available as possible for a specific protein family, and introduces new geometric requirements towards a sufficient condition for structural conservation. A region in two or more proteins is conserved when all C $\alpha$ -C $\alpha$  separations of its residues are below a certain threshold distance, and certain curvature and torsion requirements are met.

Regions that can be modeled by a mosaic of structurally conserved clusters cannot be modeled by structurally conserved regions due to the requirement that all templates must be present within a structurally conserved region.

**Complementary Software**

- Biopolymer™ for building, analyzing and visualizing macromolecular three-dimensional structures
- MOLCAD™ for visualizing surface features and physical properties essential for molecular recognition
- Surfex-Dock™ for receptor-based virtual screening and molecular docking
- RACHEL™ for *de-novo* ligand design

**Model Structurally Variable Regions<sup>5</sup>**

In comparative protein modeling, target sequences often have residues inserted relative to the template structures or have regions that are structurally different from the corresponding regions in the templates. The accuracy of modeling these variable regions can be a major factor in determining the usefulness of comparative models for computational receptor based applications.

Advanced Protein Modeling combines a knowledge based approach with an *ab initio* approach to construct structurally variable regions. The program selects from a database of protein structure fragments with environmentally constrained substitution tables and other rule based filters. The independent results from both methods are clustered to make a consensus prediction which must pass a set of rule-based filters.

**Model Sidechains<sup>6</sup>**

Advanced Protein Modeling uses a sequential strategy to predict first the backbone conformation and then the sidechain conformations. It uses a knowledge-based approach which uses torsional borrowing to restrict the rotamer conformations of specific residues, and then enumerates remaining sidechain conformations taken from a rotamer library which can be filtered by probability. The packing of these sidechains is optimized by an energy function.

**Integration**

Advanced Protein Modeling is seamlessly integrated into SYBYL® 7.3. The graphical user interface supports an intuitive workflow and offers full control of each specific aspect of the model building process.

**Hardware and Software Requirements**

Advanced Protein Modeling requires a separate license in addition to Biopolymer, and is accessible from SYBYL as well as from the system command line. SYBYL, Biopolymer and Advanced Protein Modeling run on workstations operating under IRIX® (SGI®) or Linux® (x86).

**Acknowledgements**

Advanced Protein Modeling is based on the FUGUE™ and ORCHESTRAR™ technologies, developed by Professor Sir Tom Blundell, Ph.D., and his research team at the University of Cambridge, UK.

**Validation**

FUGUE has been extensively tested in CAFASP2 (<http://cafasp.bioinfo.pl>), LiveBench2 (<http://bioinfo.pl/LiveBench>), and the CASP4 & 5 competitions (<http://predictioncenter.llnl.gov/>). It has also been used for identifying a novel super family of enzymes that catalyze the modification of guanidino groups<sup>7</sup>, and to study the evolutionary origins of the SWIB domain and the p53-binding MDM2 domain.<sup>8</sup>

ORCHESTRAR has been used in CASP7; earlier components of ORCHESTRAR have been tested in CASP4.



Ribbon model of an alignment of the proteins Coagulation Factor Xa-Trypsin Chimera, Bovine Coagulation Factor Xa, and Alpha-Thrombin Complex, created in Advanced Protein Modeling.

**References**

1. Shi, J.; Blundell, T.; Mizuguchi, K. "FUGUE: Sequence-Structure Homology Recognition Using Environment-Specific Substitution Tables and Structure-Dependent Gap Penalties." *J. Mol. Biol.* 2001, 310, 243-257.
2. Johnson, M.S.; Overington, J.P.; Blundell T.L. "Alignment and searching of common protein folds using a database of structural templates; *J.Mol.Biol.* 1993, 231, 735-752.
3. Mizuguchi, K.; Deane, C.; Blundell, T.; Overington, J. "HOMSTRAD: A Database of Protein Structure Alignments for Homologous Families." *Protein Sci.* 1998, 7, 2469-2471.
4. Montalvao, R.W.; Smith, R.E.; Lovell, S.C., Blundell, T.L. "CHORAL: a differential geometry approach to the prediction of the cores of protein structures." *Bioinformatics* 2005, 21, 3719-3725.
5. Deane, C.M.; Blundell, T.L. "CODA: A combined algorithm for predicting the structurally variable regions of protein models." *Protein Science* 2001, 10, 599-612.
6. Smith, R.E.; Lovell, S.C.; Montalvao, R.W.; Blundell, T.L. "ANDANTE." Unpublished data.
7. Shirai, H.; Blundell, T.; Mizuguchi, K. "A Novel Superfamily of Enzymes that Catalyze the Modification of Guanidino Groups." *Trends in Biochemical Sciences* 2001, 26, 465-468.
8. Bennett-Lovsey, R.; Hart, S.E.; Shirai, H.; Mizuguchi, K. "The SWIB and the MDM2 Domains are Homologous and Share a Common Fold." *Bioinformatics* 2002, 18, 620-630.



WWW.TRIPOS.COM				CONTACT_US@TRIPPOS.COM		
AUSTRALIA +61 (7) 5439 9775	CANADA +1 450 4334500	FRANCE +33 1 69 59 29 49	GERMANY +49 89 45 10 300	JAPAN +81 3 5166 1721	UNITED KINGDOM +44 1 908 650000	UNITED STATES 800-323-2960 +1-314-647-1099
©Tripos, Inc. 2006 All rights reserved. Printed in USA All trademarks are the property of their respective owners.						