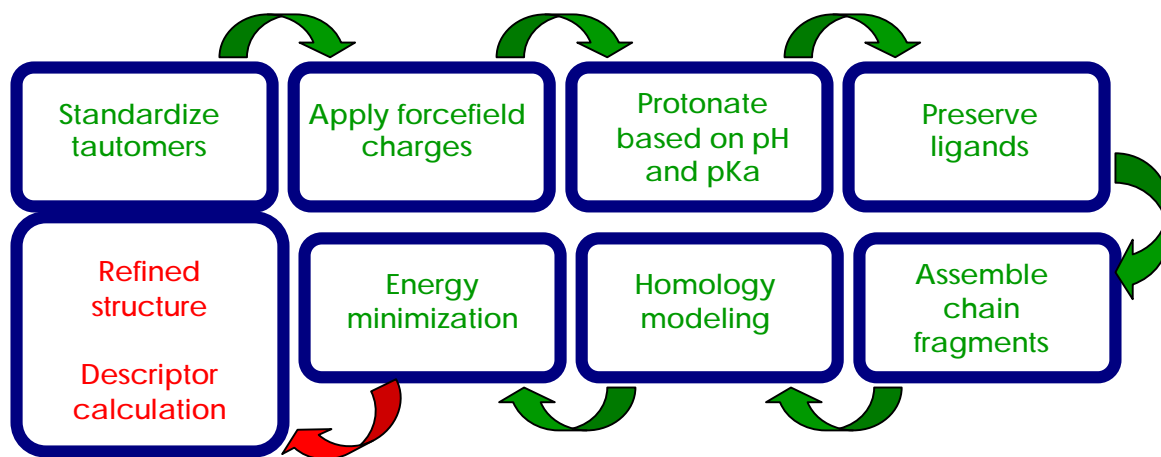


RECCR WebPDB pdb pre-processing web tool

As there is a constant search for molecular targets for small molecules, having accurate structures of target proteins would be highly beneficial. There are many x-ray crystallography- and NMR-derived structures. Unfortunately, most of these structures are error-ridden and often incomplete. Many of the more notable errors include incorrect tautomers due to lack of assigned bond order (Todorov 2006), no assignment of partial charge (Rafi 2006), missing hydrogens ((Kmuniček 2001), and because of the lack of hydrogens and partial charge, no ionization (Vogel 2006). Missing portions of proteins tend to be highly flexible, and therefore not crystallizable (Aronson 1997). Though flexible, these C-terminal, N-terminal, and loop sections of proteins are often critical binding sites (Ceci 2003, Ko 2001, Tomoo 2002). Recovering those regions of the protein that are missing can be accomplished through self-homology modeling (Aronson 1997, Singh 2006). The rest of the major errors can be repaired utilizing Reduce (Word 1999), for the tautomers, PROPKA for the appropriate protonation and ionization (Li 2005), and many functionalities within MOE (Molecular Operating Environment, Chemical Computing Group, Inc.)

To accomplish this, we created a script called WebPDB, which repairs many of the errors in PDBs, and then uses homology modeling to replace the missing regions. After using forcefields to energy minimize sidechains and complete the refining process (Rudrabhatla 2004), descriptors are calculated based on properties of the protein. These descriptors and the refined structure are available for the user for application with docking, scoring, and shape matching methods, as well as machine learning.

As pH considerations have been incorporated in the protonation and deprotonation of the structures refined by this script, the descriptors computed for the surfaces created by WebPDB will be pH-dependent,. The descriptors to be implemented in this script would include: TAE RECON (Song 2002, Whitehead 2003), PEST (Breneman 2003), and hydration based descriptors (Garcia 1997, Garde 1996).



WebPDB utilizes Reduce to correct tautomers and add initial hydrogens to the structure (Word 1999). Once hydrogens have been added, charge should be estimated on each atom using semi-empirical quantum mechanics calculations before any ionization can occur (Rafi 2006). An SVL script utilizing PROPKA is then used to properly protonate and deprotonate every ionizable atom based on the pH of the environment and the estimated pKa of that atom (Li 2005). Once these corrections have been made, the heteroatoms are removed and preserved for later re-introduction after homology modeling. Any waters in the crystal structure are then removed for consistency between structures. Once there are no heteroatoms in the model, if the chains have been fragmented due to missing loops, fragments of loops are concatenated before homology modeling can occur. After the fragments have been placed in their appropriate chains, the sequence of each chain is aligned with the structural residues so that any missing portions, such as end gaps or loops, are

prepared for homology modeling (Aronson 1997, Singh 2006). Homology modeling is performed, creating ten potential structures and keeping the average of all ten as the final model. This averaged structure, along with the heteroatoms, is energy minimized using a forcefield so that the contacts between chains of the protein, and between the heteroatoms and the protein, are optimized to give a closer approximation to the protein's structure (Fahmy 2002). After energy minimization, a triangulated surface is calculated. Descriptors can be calculated and mapped onto this surface based on the properties of the protein at those points.

There are a number of other programs that refine PDB structures:

Program	SWISS-PDB ¹	PDB2PQR ²	Protein Preparation Wizard ³	3D-JIGSAW ⁴	X3M ⁵	PROCHECK ^{6*} PROVE ^{7*} WHAT IF ^{8*}	PrimeX ⁹
Heteroatoms	N	Y	Y	N	N	Y	N
Tautomers	N	N	Y	N	N	Y	N
Protonation/ Ionization	N	Y	Y	N	N	N	N
Loop replacement	Y	N	N	Y	Y	N	Y
Homology modeling	Y	N	N	Y	Y	N	N
Energy minimization	Y	Y	Y	Y	Y	N	Y
Output format	PDB	PQR (like PDB)	Structure file compatible with Glide, CombiGlide, QSite, Liason.	PDB	NS	Analysis of structure (text)	NS

*Also known as Biotech validation Suite, ¹Schwede 2003, ²Dolinsky 2007, ³<http://www.schrodinger.com>, ⁴Bates 2001, ⁵Lund 2002, ⁶Laskowsky 1993, ⁷Pontious 1996, ⁸Vriend 1990, ⁹<http://www.schrodinger.com>, NS not specified.

References

- Aronson, N.N.; Blanchard, C.J.; Madura, J.D. Homology Modeling of Glycosyl Hydrolase Family 18 Enzymes and Proteins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*(6), 999-1005.
- Bates, P.A.; Kelley, L.A.; MacCallum, R.M.; Sternberg, M.J.E. Enhancement of Protein Modelling by Human Intervention in Applying the Automatic Programs 3D-JIGSAW and 3D-PSSM. *Proteins: Struct., Funct., and Genet., Suppl.* **2001**, *5*, 39-46.
- Breneman, C.M.; Sundling, C.M.; Sukumar, N.; Shen, L.; Katt, W.P.; Embrechts, M.J. New developments in PEST shape/property hybrid descriptors. *J. Comput.-Aided Mol. Des.* **2003**, *17*(2-4), 231-240.
- Ceci, P.; Ilari, A.; Falvo, E.; Chiancone, E. The Dps protein of *Agrobacterium tumefaciens* Does not Bind to DNA but Protects It toward Oxidative Cleavage. *J. Biol. Chem.* **2003**, *278*(22), 20319-20326.
- Fahmy, A.; Wagner, G. TreeDock: A Tool for Protein Docking Based on Minimizing van der Waals Energies. *J. Am. Chem. Soc.* **2002**, *124*(7), 1241-1250.
- Kmuniček, J.; Luengo, S.; Gago, F.; Ortiz, A.R.; Wade, R.C.; Damborský, J. Comparative Binding Energy Analysis of the Substrate Specificity of Haloalkane Dehalogenase from *Xanthobacter autotrophicus* GJ10. *Biochemistry* **2001**, *40*(30), 8905-8917.
- Ko, T.-P.; Chen, Y.-K.; Robinson, H.; Tsai, P.-C.; Gao, Y.-G.; Chen, A. P.-C.; Wang, A.H.-J.; Liang, P.-H. Mechanism of Product Chain Length Determination and the Role of a Flexible Loop in *Escherichia coli* Undecaprenyl-pyrophosphate Synthase Catalysis. *J. Biol. Chem.* **2001**, *276*(50), 47474-47482.

- Li, H.; Robertson, A.D.; Jensen, J.H. Very Fast Empirical Prediction and Rationalization of Protein pK_a Values. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*(4), 704-721.
- Lund, O.; Nielsen, M.; Lundegaard, C.; Worning, P. 2006. CPHmodels 2.0: X3M a Computer Program to Extract 3D Models. Abstract at the CASP5 conference, A102.
- Pittman, J.; Sacks, J.; Young, S.S. The Construction and Assessment of a Statistical Model for the Prediction of Protein Assay Data. *J. Chem. Inf. Comput. Sci.* **2002**, *42*(3), 729-741.
- Pontius, J.; Richelle, J.; Wodak, S.J. Quality assessment of protein 3D structures using standard atomic volumes. *J. Mol. Biol.* **1996**, *264*(1), 121-136.
- Rafi, S.B.; Cui, G.; Song, K.; Cheng, X.; Tonge, P.J.; Simmerling, C. Insight through Molecular Mechanics Poisson-Boltzmann Surface Area Calculations into the Binding Affinity of Triclosan and Three Analogues for FabI, the *E. coli* Enoyl Reductase. *J. Med. Chem.* **2006**, *49*(15), 4574-4580.
- Rudrabhatla, P.; Rajasekharan, R. Functional Characterization of Peanut Serine/Threonine/Tyrosine Protein Kinase: Molecular Docking and Inhibition Kinetics with Tyrosine Kinase Inhibitors. *Biochemistry* **2004**, *43*(38), 12123-12132.
- Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M.C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **2003**, *31*(3), 3381-3385.
- Singh, N.; Chev e, G.; Avery, M.A.; McCurdy, C.R. Comparative Protein Modeling of 1-Deoxy-D-xylulose-5-phosphate Reductoisomerase Enzyme from *Plasmodium falciparum*: A Potential Target for Antimalarial Drug Discovery. *J. Chem. Inf. Model.* **2006**, *46*(3), 1360-1370.
- Song, M.; Breneman, C.; Bi, J.; Sukumar, N.; Bennett, K.P.; Cramer, S.; Tugcu, N. Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*(6), 1347-1357.
- Todorov, N.P.; Monthoux, P.H.; Alberts, I.L. The Influence of Variations of Ligand Protonation and Tautomerism on Protein-Ligand Recognition and Binding Energy Landscape. *J. Chem. Inf. Model.* **2006**, *46*(3), 1134-1142.
- Tomoo, K.; Shen, X.; Okabe, K.; Nozoe, Y.; Fukuhara, S.; Morino, S.; Ishida, T.; Taniguchi, T.; Hasegawa, H.; Terashima, A.; Sasaki, M.; Katsuya, Y.; Kitamura, K.; Miyoshi, H.; Ishikawa, M.; Miura, K.-i. Crystal structures of 7-methylguanosine 5'-triphosphate (m^7GTP)- and P^1 -7-methylguanosine- P^3 -adenosine-5',5'-triphosphate (m^7GpppA)-bound human full-length eukaryotic initiation factor 4E: biological importance of the C-terminal flexible region. *Biochem. J.* **2002**, *362*(3), 539-544.
- Vogel, R.; Siebert, F.; Yan, E.C.Y.; Sakmar, T.P.; Hirshfeld, A.; Sheves, M. Modulating Rhodopsin Receptor Activation by Altering the pK_a of the Retinal Schiff Base. *J. Am. Chem. Soc.* **2006**, *128*(32), 10503-10512.
- Vriend, G. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **1990**, *8*(1), 52-56.
- Whitehead, C.E.; Sukumar, N.; Breneman, C.M.; Ryan, M.D. Transferable Atom Equivalent Multi-Centered Multipole Expansion Method. *J. Comp. Chem.* **2003**, *24*(4), 512-529.
- Word, J.M.; Lovell, S.C.; Richardson, J.S.; Richardson, D.C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Bio.* **1999**, *285*(4), 1735-1747.
- Zhou, Y.-H.; Zheng, Q.-C.; Li, Z.-S.; Zhang, Y.; Sun, M.; Sun, C.-C.; Si, D.; Cai, L.; Guo, Y.; Zhou, H. On the human CYP2C9*13 variant activity reduction: a molecular dynamics simulation and docking study. *Biochimie* **2006**, *88*(10), 1457-1465.