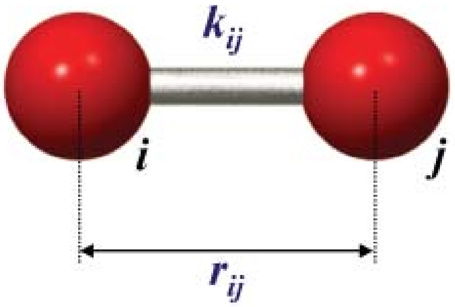
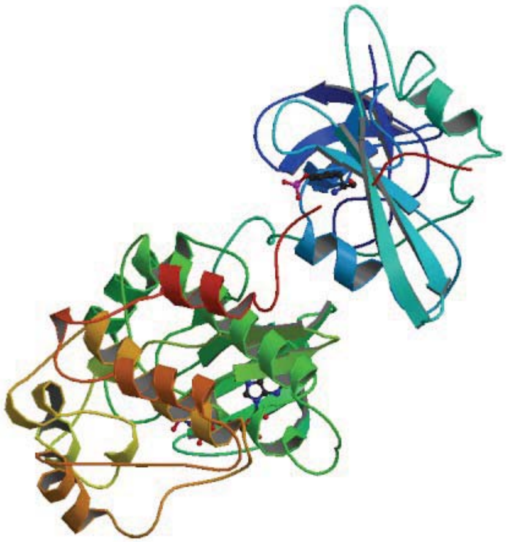
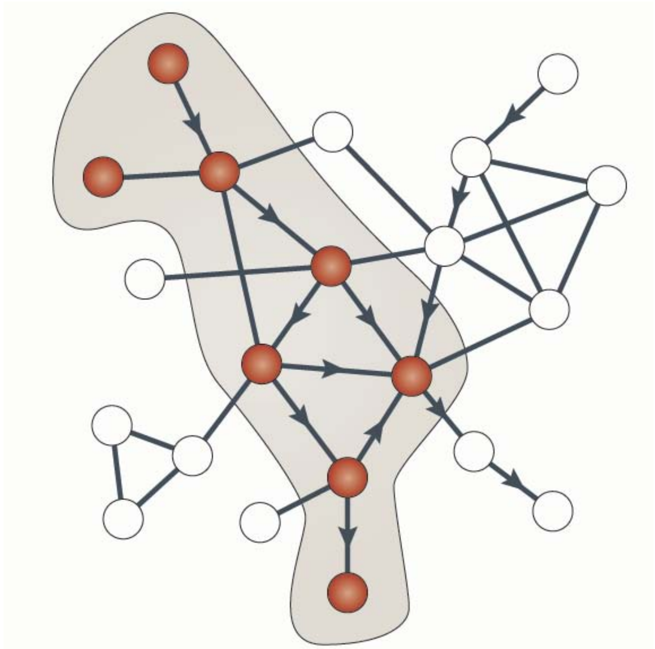
Modeling Scales



$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2,$$



Courtesy of Wenqing Xu et al. and RCSB Protein Data Bank. Used with permission.



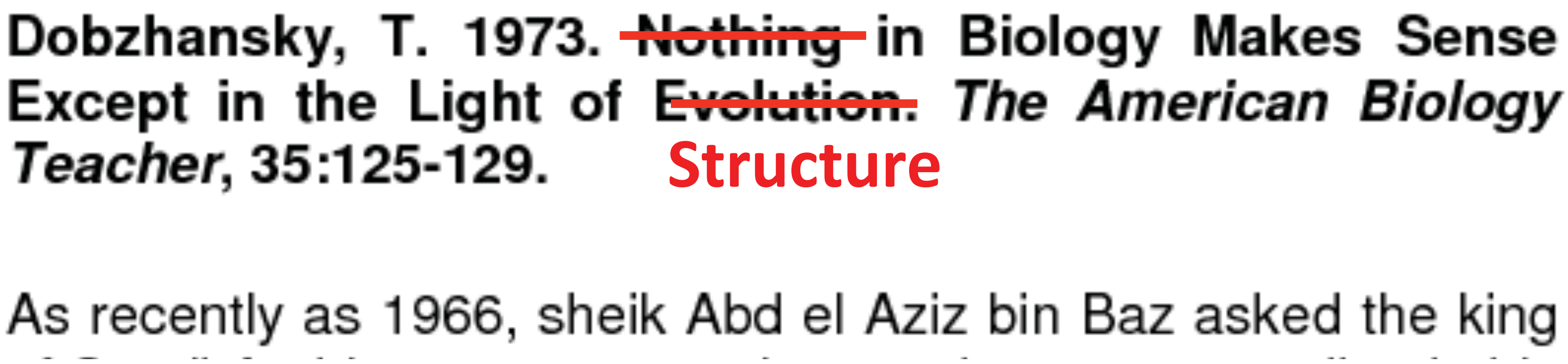
Courtesy of Macmillan Publishers Limited. Used with permission. Source: Barabási, Albert-László, Natali Gulbahce, et al. "Network Medicine: A Network-based Approach to Human Disease." *Nature Reviews Genetics* 12, no. 1 (2011): 56-68.

Atom **Protein** **Network**

- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models

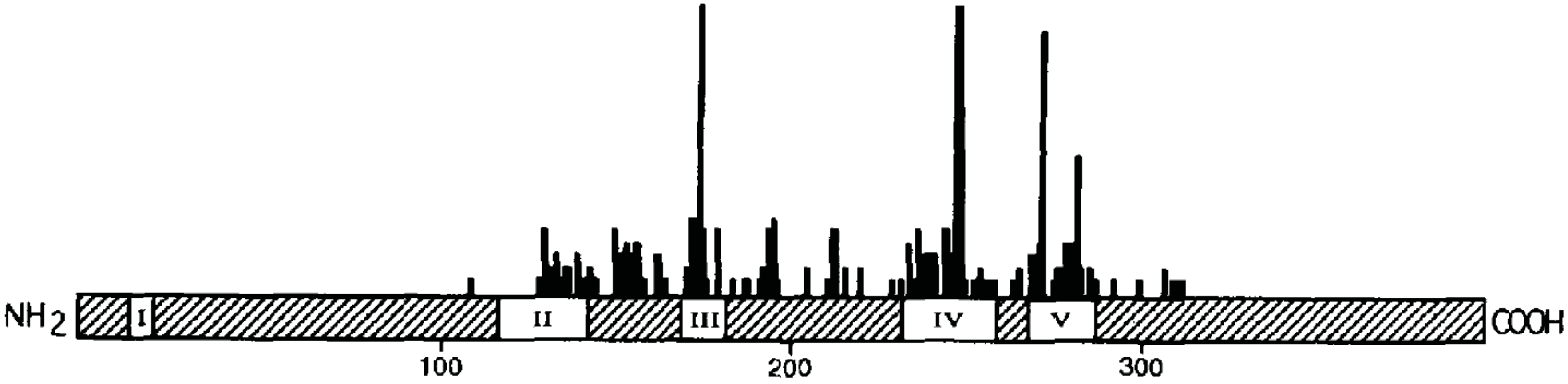
Lecture 12

Introduction to protein structure

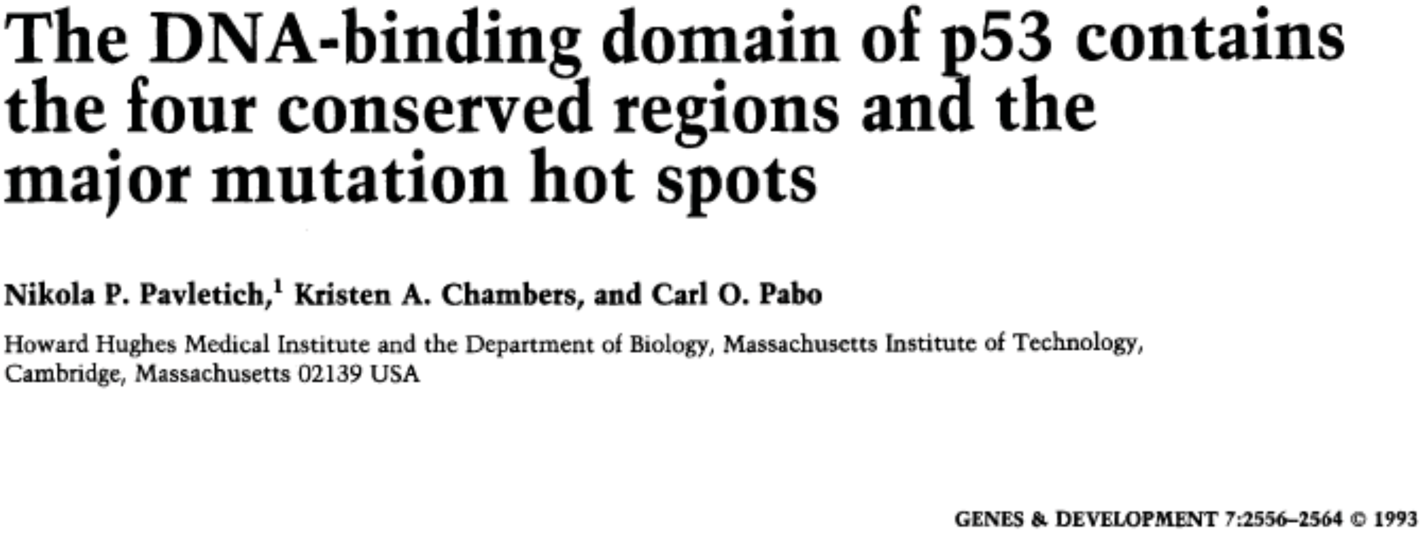


Dobzhansky, T. 1973. ~~Nothing~~ Little in Biology Makes Sense Except in the Light of ~~Evolution.~~ Structure *The American Biology Teacher*, 35:125-129.

As recently as 1966, sheik Abd el Aziz bin Baz asked the king



Source: Pavletich, Nikola P., Kristen A. Chambers and Carl O. Pabo. "The DNA-binding Domain of p53 Contains the Four Conserved Regions and the Major Mutation Hot Spots." *Genes & Development* 7, no. 12b (1993): 2556-64.

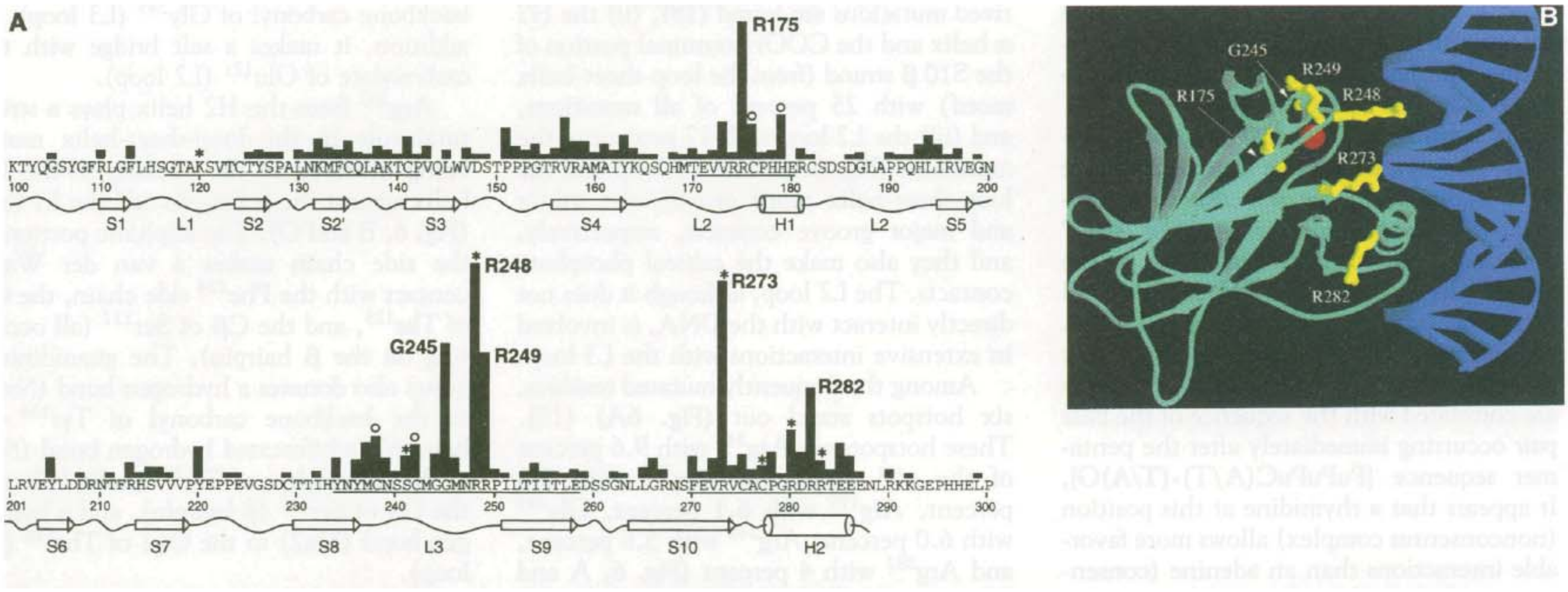


The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots

Nikola P. Pavletich,[1] Kristen A. Chambers, and Carl O. Pabo

Howard Hughes Medical Institute and the Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

GENES & DEVELOPMENT 7:2556–2564 © 1993



Source: Cho, Yunje, Svetlana Gorina, et al. "Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations." *Science* 265, no. 5170 (1994): 346-55.

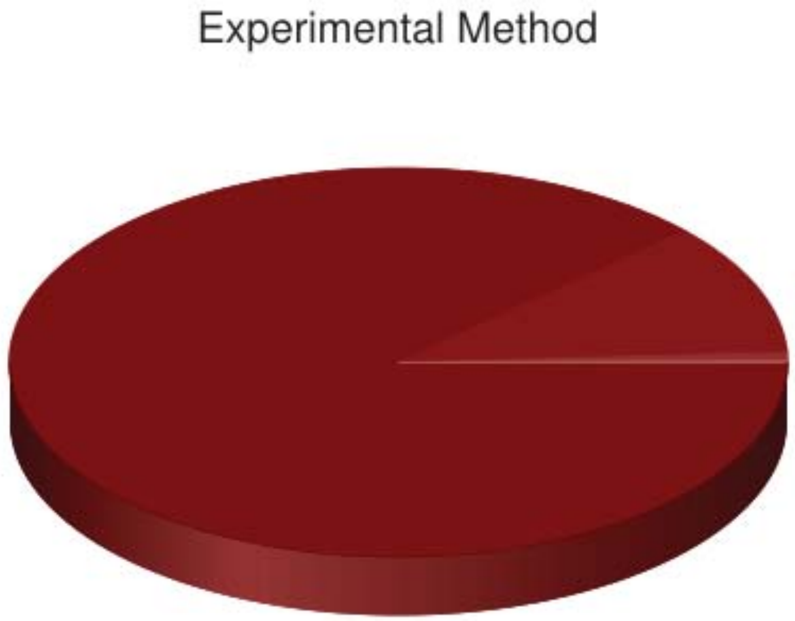
Fig. 6. The residues most frequently mutated in cancer are at or near the protein-DNA interface. (**A**) Sequence of the p53 core domain showing the conserved regions (underlined), and the secondary structure elements. The number of tumor-derived missense mutations at each residue are indicated by the bar graph and the six most frequently mutated residues are labeled (*18*). Residues involved in DNA binding are indicated by asterisks, and those involved in binding the zinc atom are indicated by circles. Single letter abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. (**B**) Ribbon drawing of the p53 core domain–DNA complex showing the six most frequently mutated residues of p53. The side chains of these residues are colored yellow, the core domain is light blue, and the DNA is dark blue. The zinc atom is shown as a red sphere.

Crystal Structure of a p53 Tumor Suppressor–DNA Complex: Understanding Tumorigenic Mutations

Yunje Cho, Svetlana Gorina, Philip D. Jeffrey, Nikola P. Pavletich

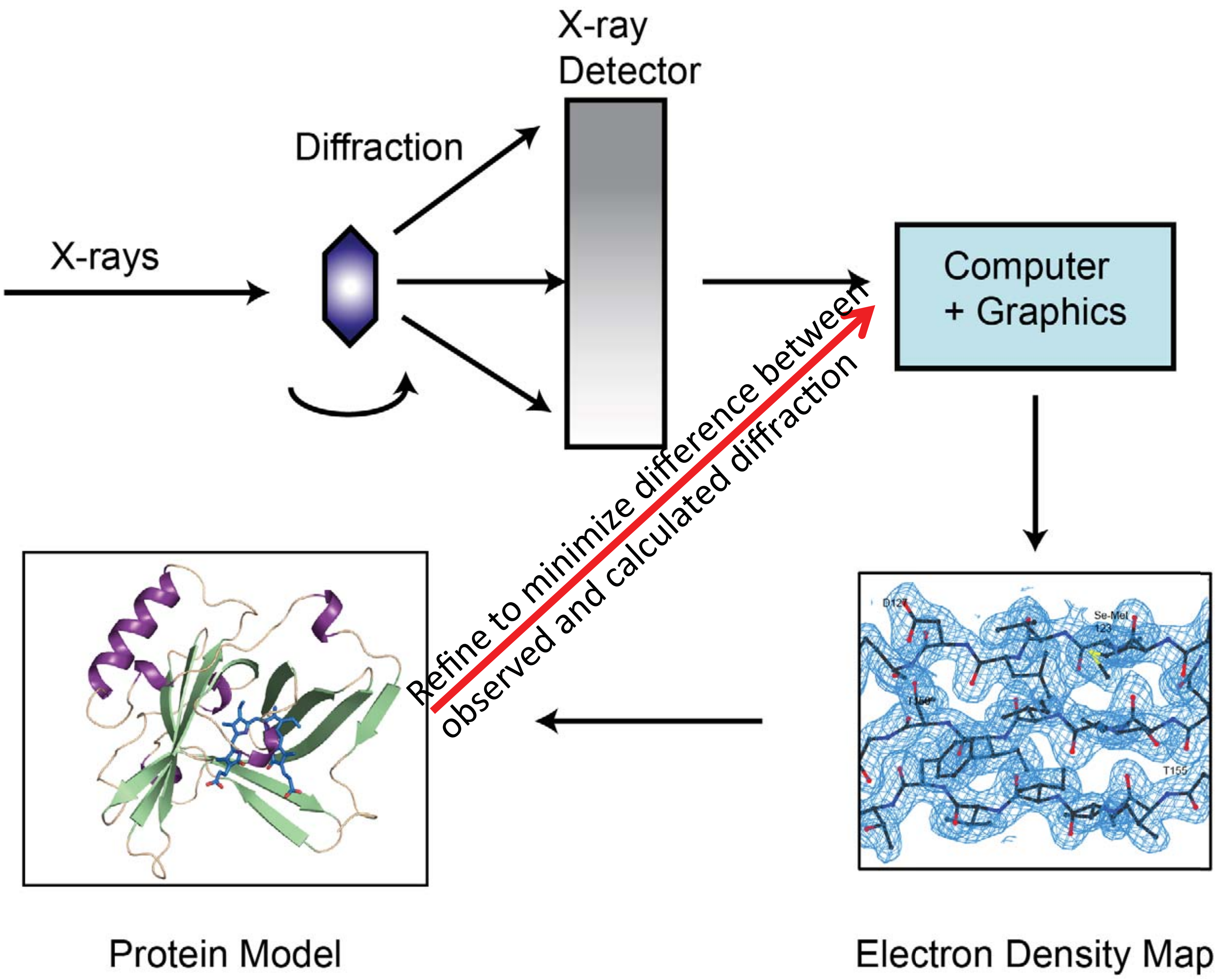
SCIENCE • VOL. 265 • 15 JULY 1994

http://www.rcsb.org/pdb



X-ray (78934)
Solution NMR (9828)
Electron Microscopy (522)
Solid-State NMR (56)
Hybrid (52)
Neutron Diffraction (43)
Fiber Diffraction (37)
Electron Crystallography (34)
Solution Scattering (32)
Other (23)

Overview of the X-ray Crystallographic Method



X-rays

Diffraction

X-ray Detector

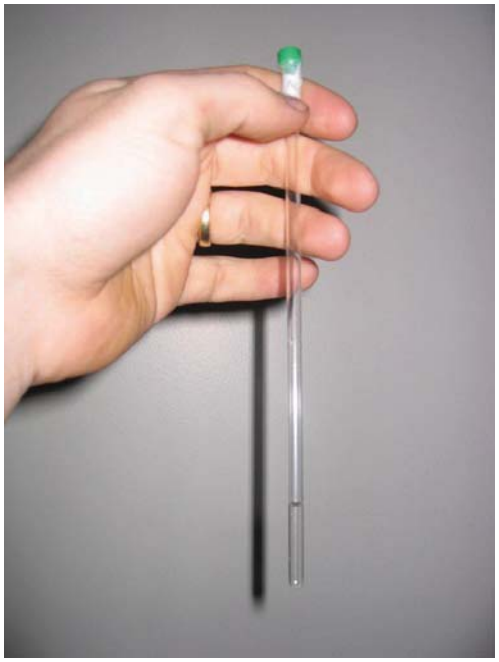
Computer + Graphics

Refine to minimize difference between observed and calculated diffraction

Protein Model

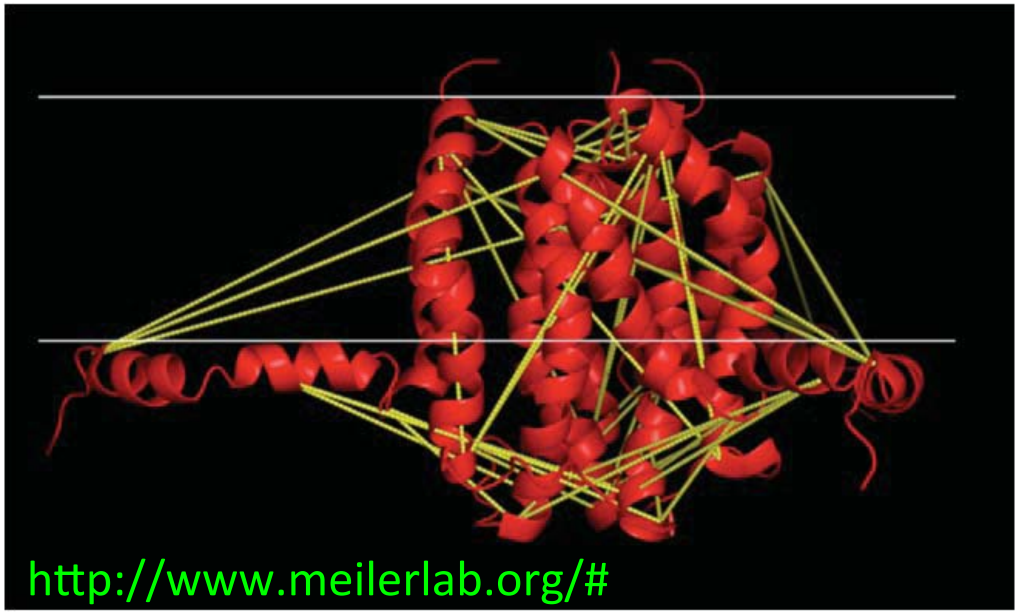
Electron Density Map

NMR



Courtesy of Kjaergaard on wikipedia.
Photograph in the public domain.

Courtesy of MartinSaunders on wikipedia.
Photograph in the public domain.



Structure are “solved” not observed

- Both crystallography and NMR depend on computational methods to find the structure (or structures) that best agree with experimental data.

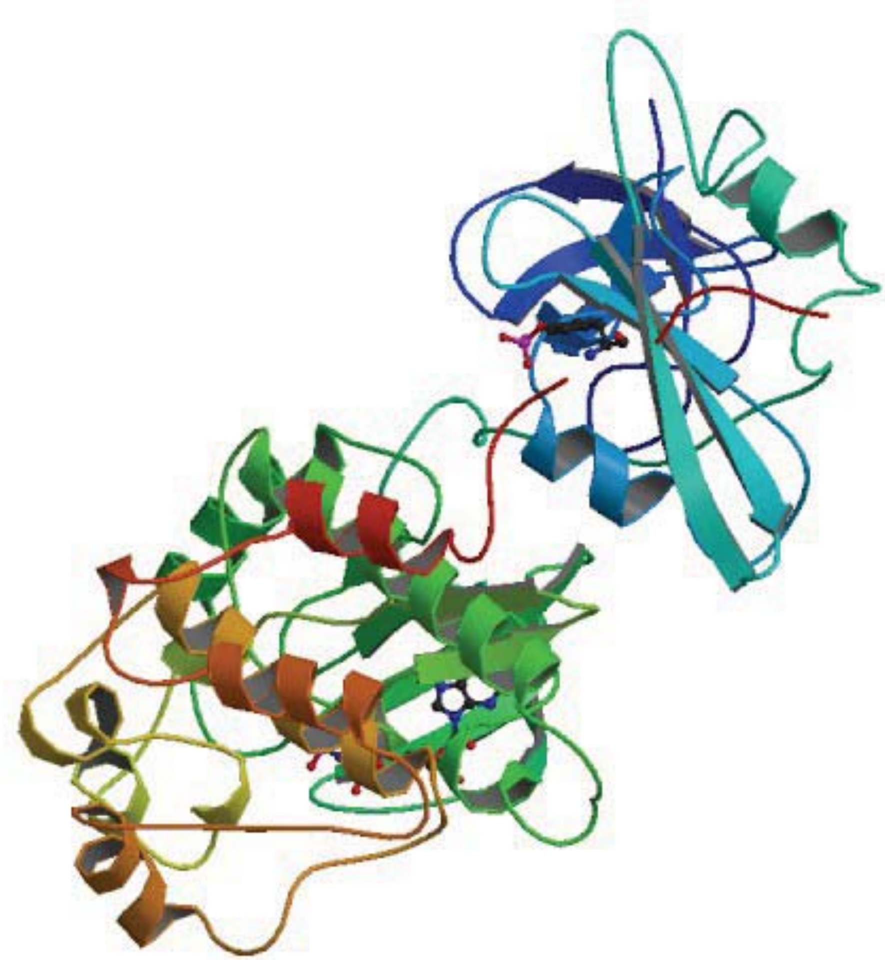
Predicting Structure

- Closely tied to the computational challenges of interpreting X-ray and NMR data
- A key topic in our lectures

Challenges of Structural Bioinformatics

courtesy of Russ Altman & Jonathan Dugan
in *Structural Bioinformatics, Philip E. Bourne & Helge Weissig, editors*

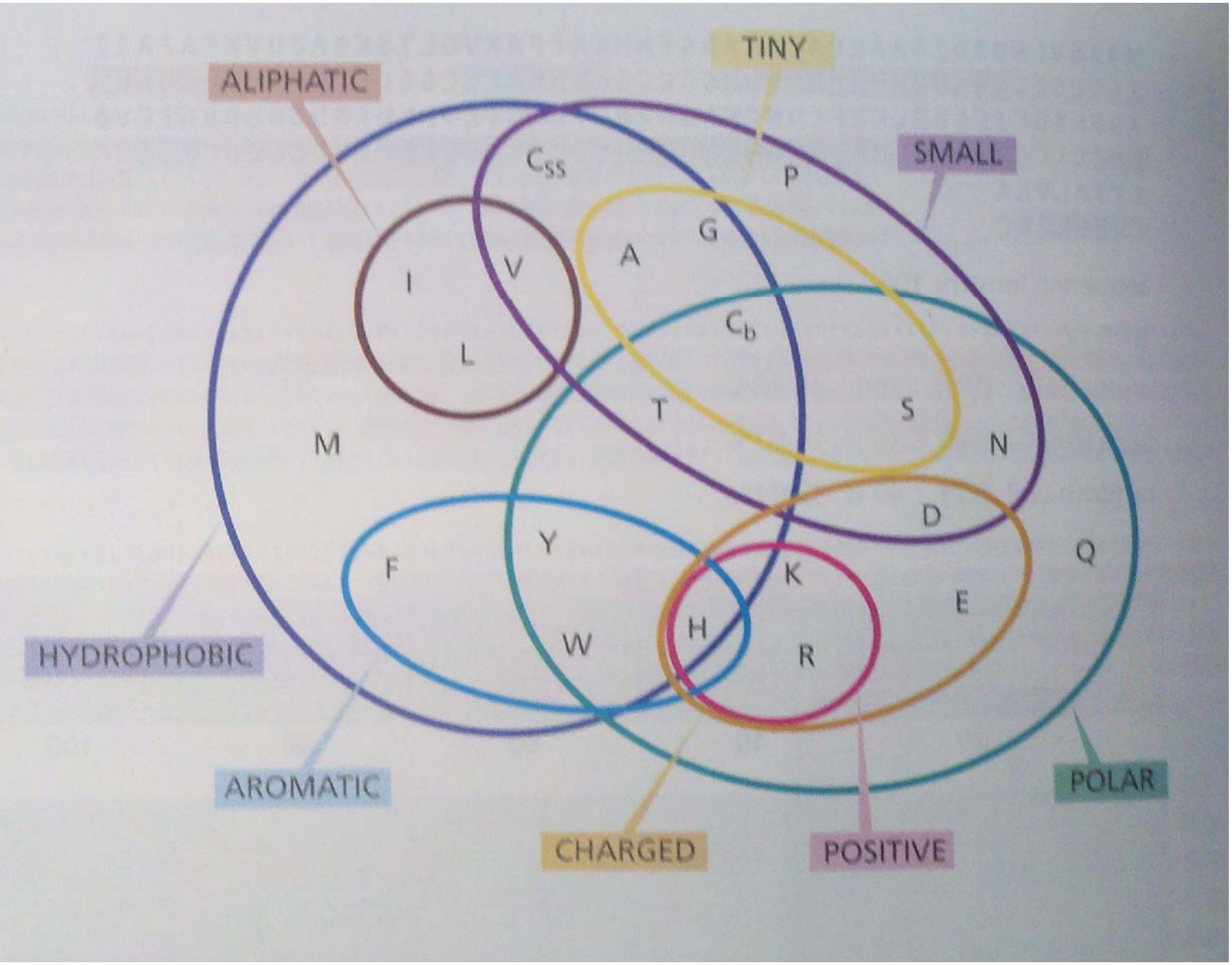
1. Structural data are not linear - can't apply string algorithms
2. Search space is continuous/infinite
3. Structure is determined by physics, in a subtle way that resists simplification
4. Human vs. computer interfaces to structure (visualization vs. coordinates) are very different
5. Experimental structural data are imperfect & incomplete
6. Proteins related in terms of structure may have very dissimilar sequences and so be hard to identify
7. We don't know much about some large classes of important proteins
8. Structural biology for the most part describes parts of a whole - assembly is tricky



Courtesy of Wenqing Xu et al. and RCSB Protein Data Bank. Used with permission.

Read posted material for details on primary, secondary, tertiary structure, alpha helices, beta sheets and more

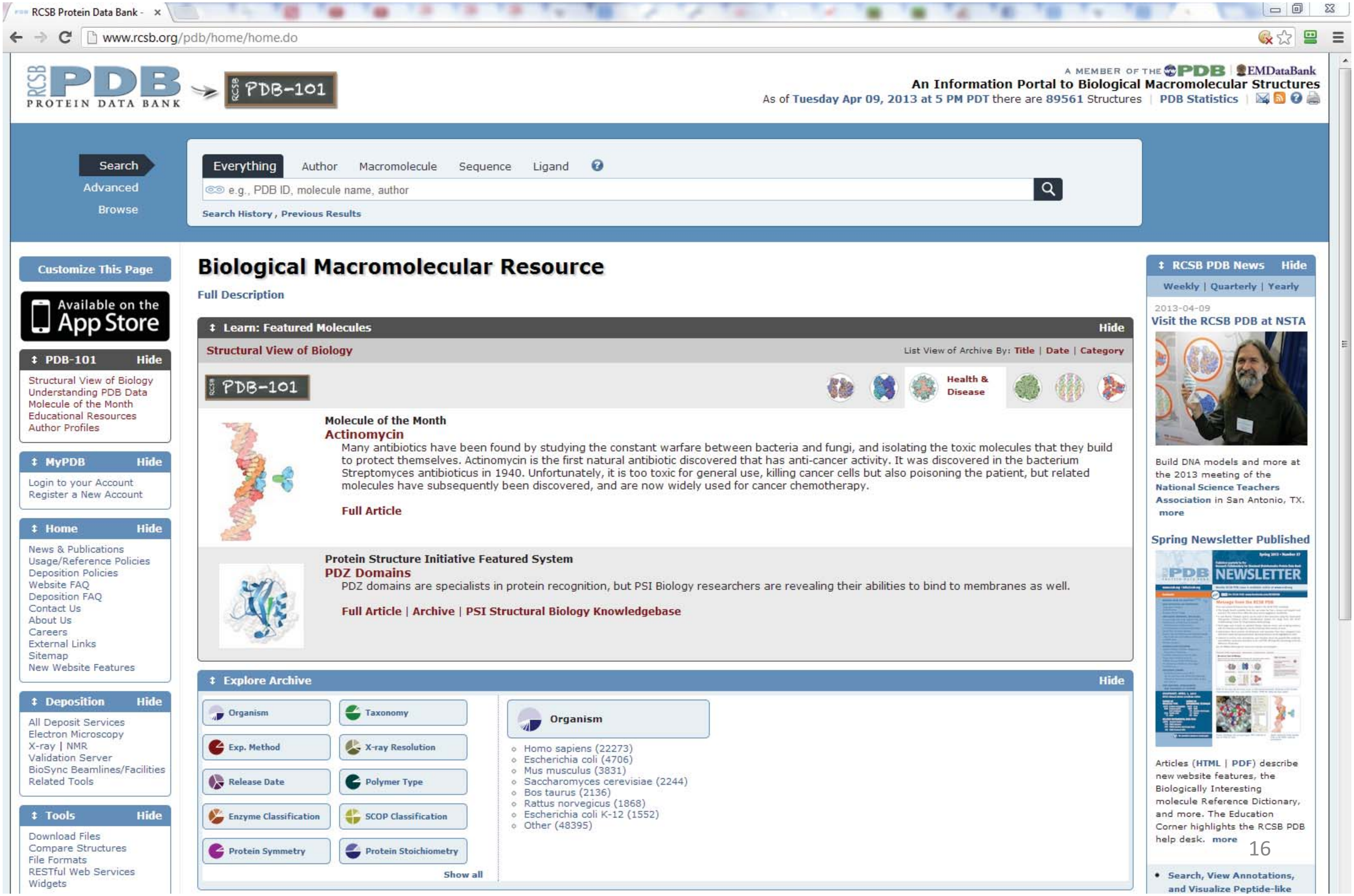
Get to know the amino acids



Source: Figure 11.18 in Zvelebil, Marketa J., and Jeremy O. Baum. "Understanding Bioinformatics." *Garland Science*, 2008.

Figure 11.18 inn Zvelebil and Baum

http://www.rcsb.org/pdb



RCSB Protein Data Bank -

www.rcsb.org/pdb/home/home.do

PROTEIN DATA BANK

PDB-101

A MEMBER OF THE PDB | EMDataBank

An Information Portal to Biological Macromolecular Structures

As of Tuesday Apr 09, 2013 at 5 PM PDT there are 89561 Structures | PDB Statistics

Search
Advanced
Browse

Everything | Author | Macromolecule | Sequence | Ligand

e.g., PDB ID, molecule name, author

Search History , Previous Results

Customize This Page

Available on the App Store

PDB-101 Hide

Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles

MyPDB Hide

Login to your Account
Register a New Account

Home Hide

News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
Sitemap
New Website Features

Deposition Hide

All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamlines/Facilities
Related Tools

Tools Hide

Download Files
Compare Structures
File Formats
RESTful Web Services
Widgets

Biological Macromolecular Resource

Full Description

Learn: Featured Molecules Hide

Structural View of Biology List View of Archive By: Title | Date | Category

PDB-101 Health & Disease

Molecule of the Month

Actinomycin

Many antibiotics have been found by studying the constant warfare between bacteria and fungi, and isolating the toxic molecules that they build to protect themselves. Actinomycin is the first natural antibiotic discovered that has anti-cancer activity. It was discovered in the bacterium Streptomyces antibioticus in 1940. Unfortunately, it is too toxic for general use, killing cancer cells but also poisoning the patient, but related molecules have subsequently been discovered, and are now widely used for cancer chemotherapy.

Full Article

Protein Structure Initiative Featured System

PDZ Domains

PDZ domains are specialists in protein recognition, but PSI Biology researchers are revealing their abilities to bind to membranes as well.

Full Article | Archive | PSI Structural Biology Knowledgebase

Explore Archive Hide

Organism | Taxonomy
Exp. Method | X-ray Resolution
Release Date | Polymer Type
Enzyme Classification | SCOP Classification
Protein Symmetry | Protein Stoichiometry
Show all

Organism

Organism
Homo sapiens (22273)
Escherichia coli (4706)
Mus musculus (3831)
Saccharomyces cerevisiae (2244)
Bos taurus (2136)
Rattus norvegicus (1868)
Escherichia coli K-12 (1552)
Other (48395)

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2013-04-09

Visit the RCSB PDB at NSTA

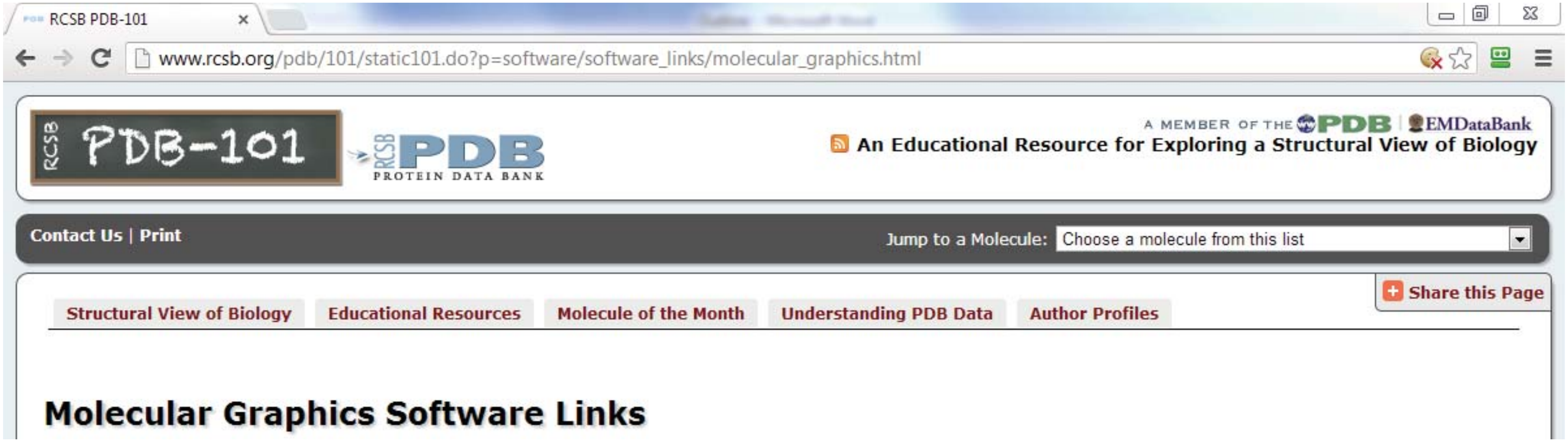
Build DNA models and more at the 2013 meeting of the National Science Teachers Association in San Antonio, TX. more

Spring Newsletter Published

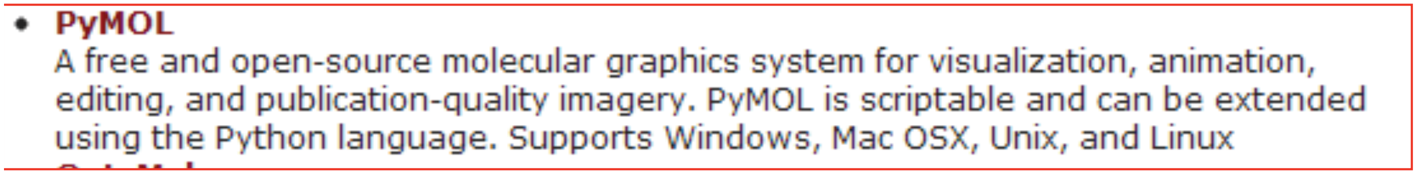
Articles (HTML | PDF) describe new website features, the Biologically Interesting molecule Reference Dictionary, and more. The Education Corner highlights the RCSB PDB help desk. more

- Search, View Annotations, and Visualize Peptide-like

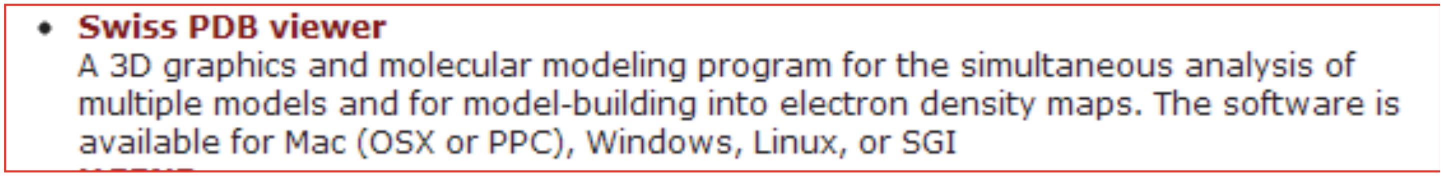
Courtesy of RCSB Protein Data Bank. Used with permission.


RCSB PDB-101
www.rcsb.org/pdb/101/static101.do?p=software/software_links/molecular_graphics.html
RCSB PDB-101
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE PDB | EMDataBank
An Educational Resource for Exploring a Structural View of Biology
Contact Us | Print
Jump to a Molecule:
Choose a molecule from this list
Share this Page
Structural View of Biology
Educational Resources
Molecule of the Month
Understanding PDB Data
Author Profiles

Molecular Graphics Software Links



- **PyMOL**
  A free and open-source molecular graphics system for visualization, animation, editing, and publication-quality imagery. PyMOL is scriptable and can be extended using the Python language. Supports Windows, Mac OSX, Unix, and Linux

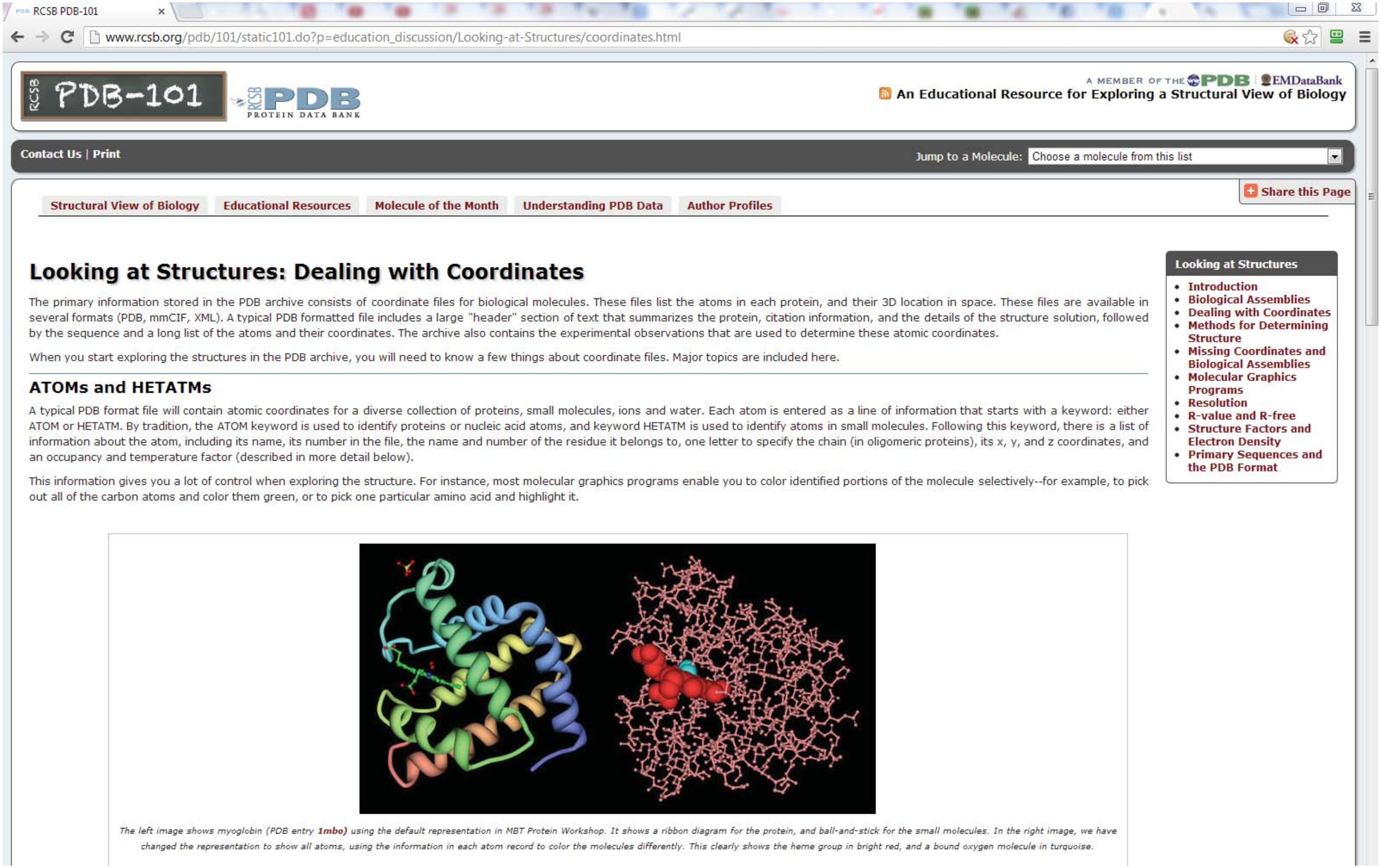


- **Swiss PDB viewer**
  A 3D graphics and molecular modeling program for the simultaneous analysis of multiple models and for model-building into electron density maps. The software is available for Mac (OSX or PPC), Windows, Linux, or SGI

Courtesy of RCSB Protein Data Bank. Used with permission.

Describing structures

- repeating elements
- x,y,z coordinates
- internal coordinates


RCSB PDB-101
www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/coordinates.html
RCSB PDB-101
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE PDB | EMDataBank
An Educational Resource for Exploring a Structural View of Biology
Contact Us | Print
Jump to a Molecule: Choose a molecule from this list
Share this Page
Structural View of Biology
Educational Resources
Molecule of the Month
Understanding PDB Data
Author Profiles

Looking at Structures: Dealing with Coordinates

The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). A typical PDB formatted file includes a large "header" section of text that summarizes the protein, citation information, and the details of the structure solution, followed by the sequence and a long list of the atoms and their coordinates. The archive also contains the experimental observations that are used to determine these atomic coordinates.

When you start exploring the structures in the PDB archive, you will need to know a few things about coordinate files. Major topics are included here.

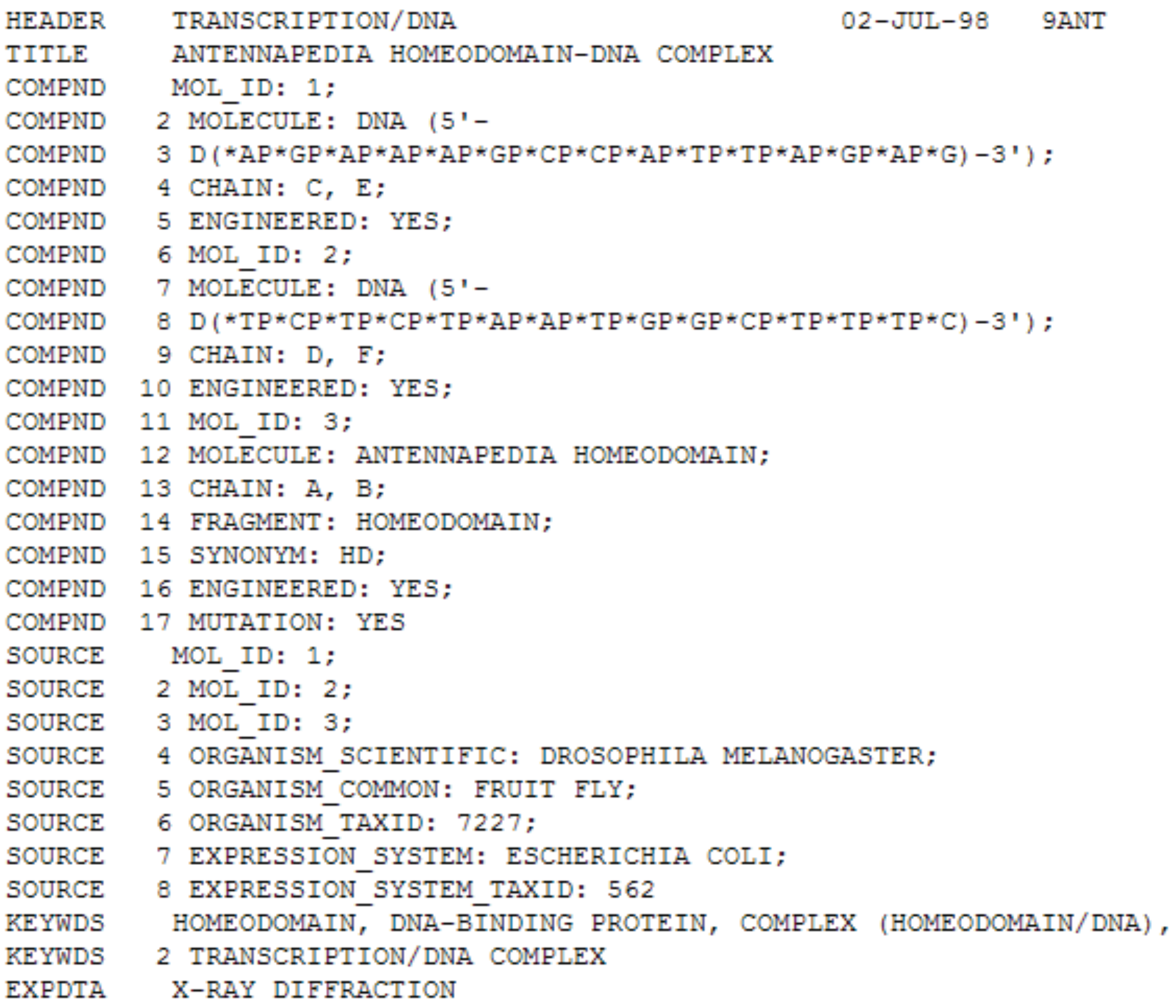
ATOMs and HETATMs

A typical PDB format file will contain atomic coordinates for a diverse collection of proteins, small molecules, ions and water. Each atom is entered as a line of information that starts with a keyword: either ATOM or HETATM. By tradition, the ATOM keyword is used to identify proteins or nucleic acid atoms, and keyword HETATM is used to identify atoms in small molecules. Following this keyword, there is a list of information about the atom, including its name, its number in the file, the name and number of the residue it belongs to, one letter to specify the chain (in oligomeric proteins), its x, y, and z coordinates, and an occupancy and temperature factor (described in more detail below).

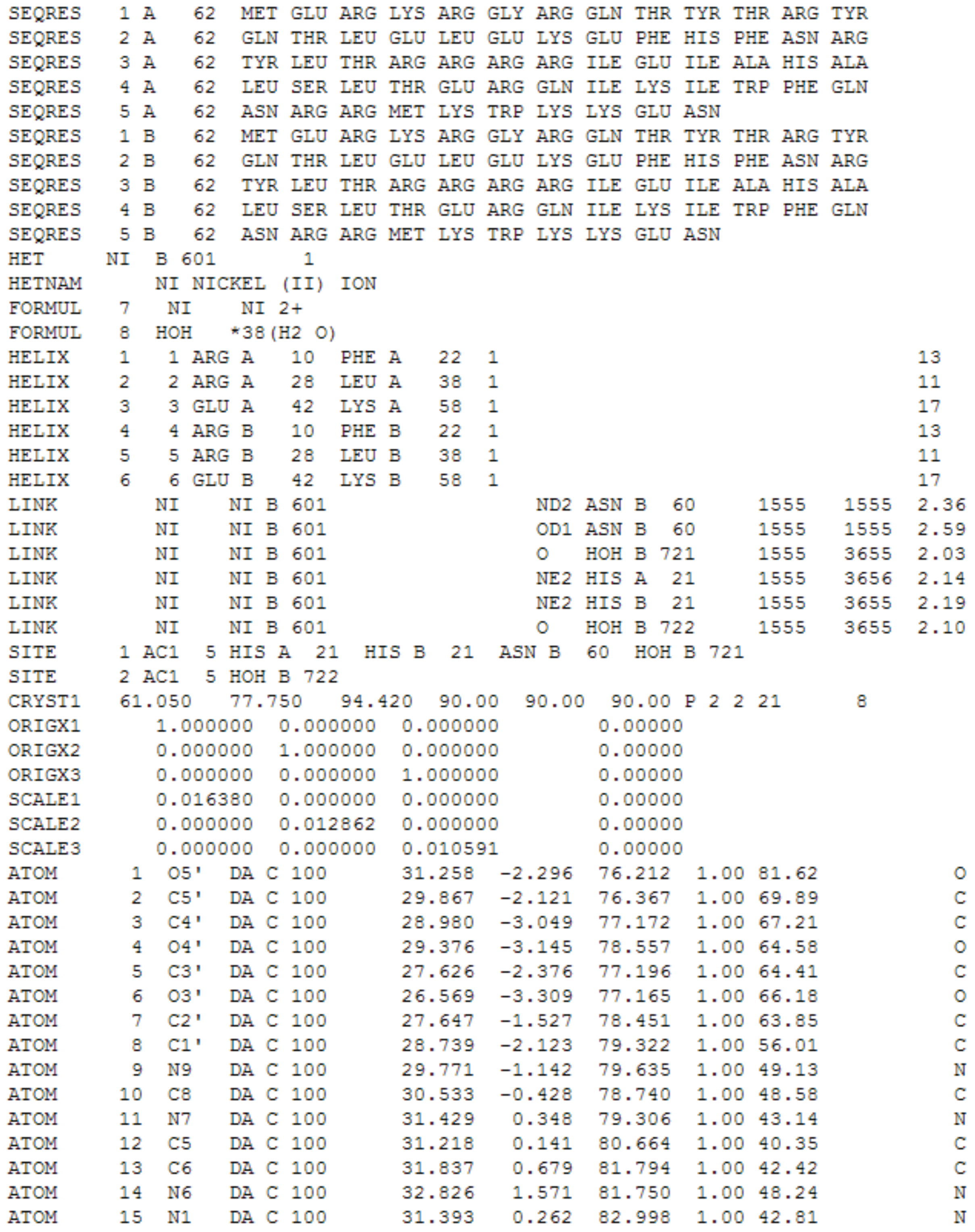
This information gives you a lot of control when exploring the structure. For instance, most molecular graphics programs enable you to color identified portions of the molecule selectively--for example, to pick out all of the carbon atoms and color them green, or to pick one particular amino acid and highlight it.

The left image shows myoglobin (PDB entry ***1mbo)*** *using the default representation in MBT Protein Workshop. It shows a ribbon diagram for the protein, and ball-and-stick for the small molecules. In the right image, we have changed the representation to show all atoms, using the information in each atom record to color the molecules differently. This clearly shows the heme group in bright red, and a bound oxygen molecule in turquoise.*

Courtesy of RCSB Protein Data Bank. Used with permission.

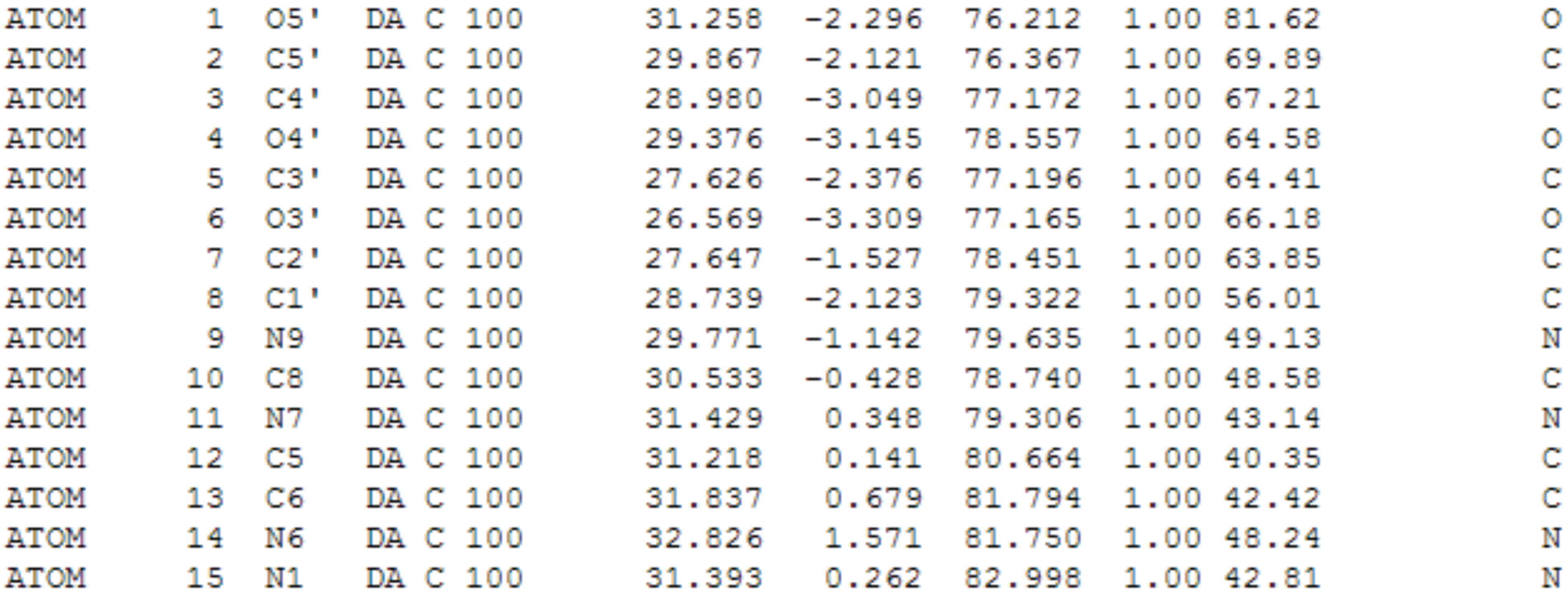


```
HEADER    TRANSCRIPTION/DNA                       02-JUL-98   9ANT
TITLE     ANTENNAPEDIA HOMEODOMAIN-DNA COMPLEX
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: DNA (5'-
COMPND   3 D(*AP*GP*AP*AP*AP*GP*CP*CP*AP*TP*TP*AP*GP*AP*G)-3');
COMPND   4 CHAIN: C, E;
COMPND   5 ENGINEERED: YES;
COMPND   6 MOL_ID: 2;
COMPND   7 MOLECULE: DNA (5'-
COMPND   8 D(*TP*CP*TP*CP*TP*AP*AP*TP*GP*GP*CP*TP*TP*TP*C)-3');
COMPND   9 CHAIN: D, F;
COMPND  10 ENGINEERED: YES;
COMPND  11 MOL_ID: 3;
COMPND  12 MOLECULE: ANTENNAPEDIA HOMEODOMAIN;
COMPND  13 CHAIN: A, B;
COMPND  14 FRAGMENT: HOMEODOMAIN;
COMPND  15 SYNONYM: HD;
COMPND  16 ENGINEERED: YES;
COMPND  17 MUTATION: YES
SOURCE    MOL_ID: 1;
SOURCE   2 MOL_ID: 2;
SOURCE   3 MOL_ID: 3;
SOURCE   4 ORGANISM_SCIENTIFIC: DROSOPHILA MELANOGASTER;
SOURCE   5 ORGANISM_COMMON: FRUIT FLY;
SOURCE   6 ORGANISM_TAXID: 7227;
SOURCE   7 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   8 EXPRESSION_SYSTEM_TAXID: 562
KEYWDS    HOMEODOMAIN, DNA-BINDING PROTEIN, COMPLEX (HOMEODOMAIN/DNA),
KEYWDS   2 TRANSCRIPTION/DNA COMPLEX
EXPDTA    X-RAY DIFFRACTION
```



```
SEQRES   1 A   62  MET GLU ARG LYS ARG GLY ARG GLN THR TYR THR ARG TYR
SEQRES   2 A   62  GLN THR LEU GLU LEU GLU LYS GLU PHE HIS PHE ASN ARG
SEQRES   3 A   62  TYR LEU THR ARG ARG ARG ARG ILE GLU ILE ALA HIS ALA
SEQRES   4 A   62  LEU SER LEU THR GLU ARG GLN ILE LYS ILE TRP PHE GLN
SEQRES   5 A   62  ASN ARG ARG MET LYS TRP LYS LYS GLU ASN
SEQRES   1 B   62  MET GLU ARG LYS ARG GLY ARG GLN THR TYR THR ARG TYR
SEQRES   2 B   62  GLN THR LEU GLU LEU GLU LYS GLU PHE HIS PHE ASN ARG
SEQRES   3 B   62  TYR LEU THR ARG ARG ARG ARG ILE GLU ILE ALA HIS ALA
SEQRES   4 B   62  LEU SER LEU THR GLU ARG GLN ILE LYS ILE TRP PHE GLN
SEQRES   5 B   62  ASN ARG ARG MET LYS TRP LYS LYS GLU ASN
HET     NI  B 601       1
HETNAM      NI NICKEL (II) ION
FORMUL   7   NI    NI 2+
FORMUL   8  HOH   *38(H2 O)
HELIX    1   1 ARG A   10  PHE A   22  1                                  13
HELIX    2   2 ARG A   28  LEU A   38  1                                  11
HELIX    3   3 GLU A   42  LYS A   58  1                                  17
HELIX    4   4 ARG B   10  PHE B   22  1                                  13
HELIX    5   5 ARG B   28  LEU B   38  1                                  11
HELIX    6   6 GLU B   42  LYS B   58  1                                  17
LINK        NI    NI B 601                 ND2 ASN B  60     1555   1555  2.36
LINK        NI    NI B 601                 OD1 ASN B  60     1555   1555  2.59
LINK        NI    NI B 601                 O   HOH B 721     1555   3655  2.03
LINK        NI    NI B 601                 NE2 HIS A  21     1555   3656  2.14
LINK        NI    NI B 601                 NE2 HIS B  21     1555   3655  2.19
LINK        NI    NI B 601                 O   HOH B 722     1555   3655  2.10
SITE     1 AC1  5 HIS A  21  HIS B  21  ASN B  60  HOH B 721
SITE     2 AC1  5 HOH B 722
CRYST1   61.050   77.750   94.420  90.00  90.00  90.00 P 2 2 21      8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.016380  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012862  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010591        0.00000
ATOM      1  O5'  DA C 100      31.258  -2.296  76.212  1.00 81.62           O
ATOM      2  C5'  DA C 100      29.867  -2.121  76.367  1.00 69.89           C
ATOM      3  C4'  DA C 100      28.980  -3.049  77.172  1.00 67.21           C
ATOM      4  O4'  DA C 100      29.376  -3.145  78.557  1.00 64.58           O
ATOM      5  C3'  DA C 100      27.626  -2.376  77.196  1.00 64.41           C
ATOM      6  O3'  DA C 100      26.569  -3.309  77.165  1.00 66.18           O
ATOM      7  C2'  DA C 100      27.647  -1.527  78.451  1.00 63.85           C
ATOM      8  C1'  DA C 100      28.739  -2.123  79.322  1.00 56.01           C
ATOM      9  N9   DA C 100      29.771  -1.142  79.635  1.00 49.13           N
ATOM     10  C8   DA C 100      30.533  -0.428  78.740  1.00 48.58           C
ATOM     11  N7   DA C 100      31.429   0.348  79.306  1.00 43.14           N
ATOM     12  C5   DA C 100      31.218   0.141  80.664  1.00 40.35           C
ATOM     13  C6   DA C 100      31.837   0.679  81.794  1.00 42.42           C
ATOM     14  N6   DA C 100      32.826   1.571  81.750  1.00 48.24           N
ATOM     15  N1   DA C 100      31.393   0.262  82.998  1.00 42.81           N
```

X Y Z Occupancy B



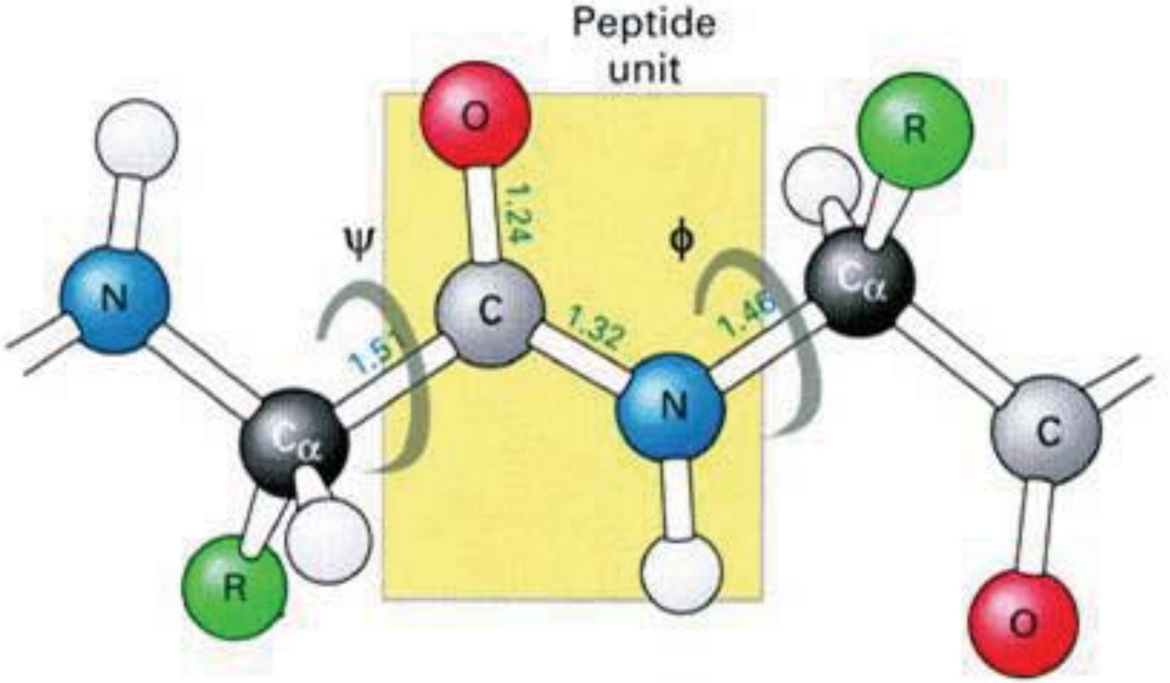
```
ATOM      1  O5'  DA C 100      31.258  -2.296  76.212  1.00 81.62           O
ATOM      2  C5'  DA C 100      29.867  -2.121  76.367  1.00 69.89           C
ATOM      3  C4'  DA C 100      28.980  -3.049  77.172  1.00 67.21           C
ATOM      4  O4'  DA C 100      29.376  -3.145  78.557  1.00 64.58           O
ATOM      5  C3'  DA C 100      27.626  -2.376  77.196  1.00 64.41           C
ATOM      6  O3'  DA C 100      26.569  -3.309  77.165  1.00 66.18           O
ATOM      7  C2'  DA C 100      27.647  -1.527  78.451  1.00 63.85           C
ATOM      8  C1'  DA C 100      28.739  -2.123  79.322  1.00 56.01           C
ATOM      9  N9   DA C 100      29.771  -1.142  79.635  1.00 49.13           N
ATOM     10  C8   DA C 100      30.533  -0.428  78.740  1.00 48.58           C
ATOM     11  N7   DA C 100      31.429   0.348  79.306  1.00 43.14           N
ATOM     12  C5   DA C 100      31.218   0.141  80.664  1.00 40.35           C
ATOM     13  C6   DA C 100      31.837   0.679  81.794  1.00 42.42           C
ATOM     14  N6   DA C 100      32.826   1.571  81.750  1.00 48.24           N
ATOM     15  N1   DA C 100      31.393   0.262  82.998  1.00 42.81           N
```

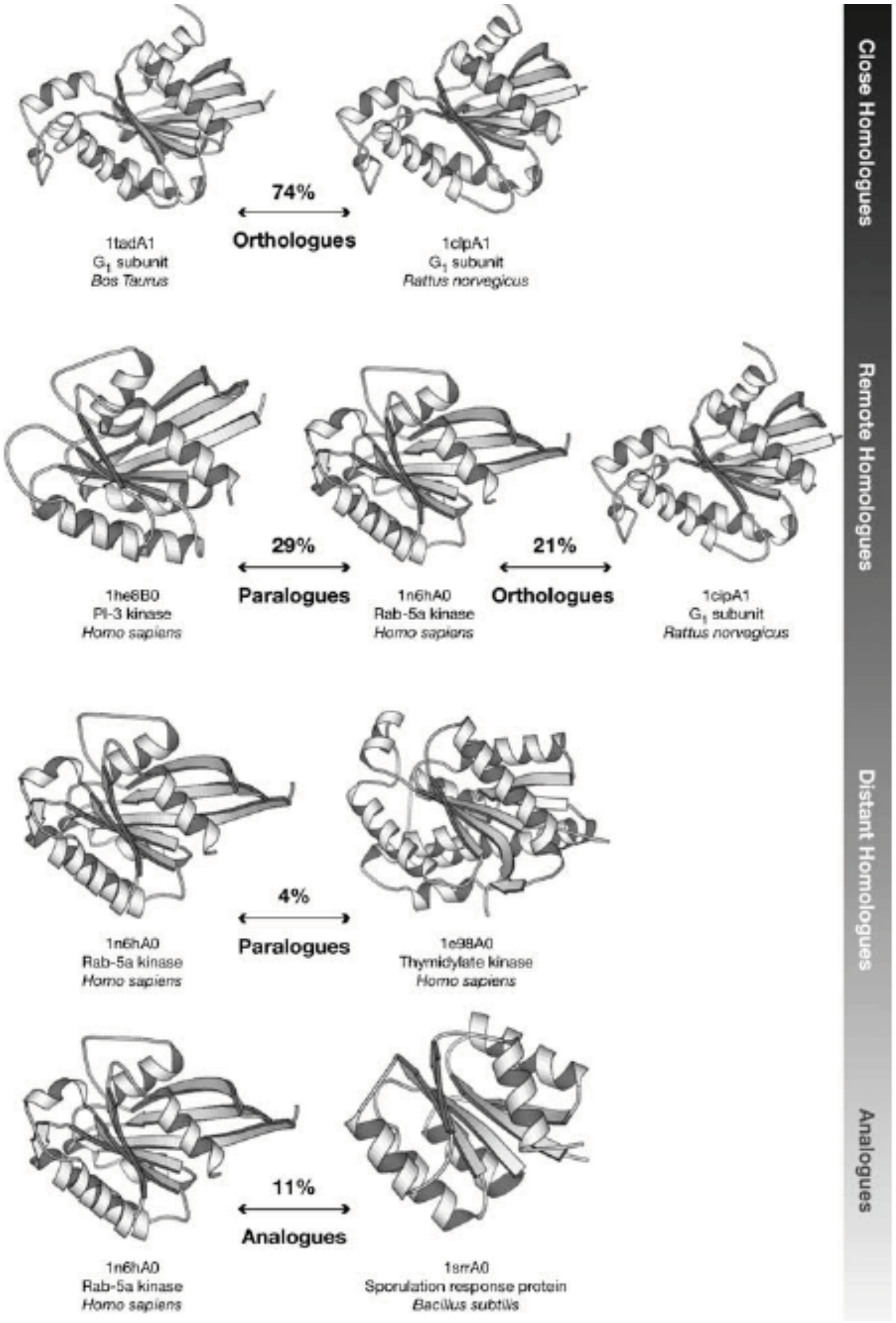
High values of B correspond to more thermal motion (range 0-100)

http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/coordinates.html

for details.

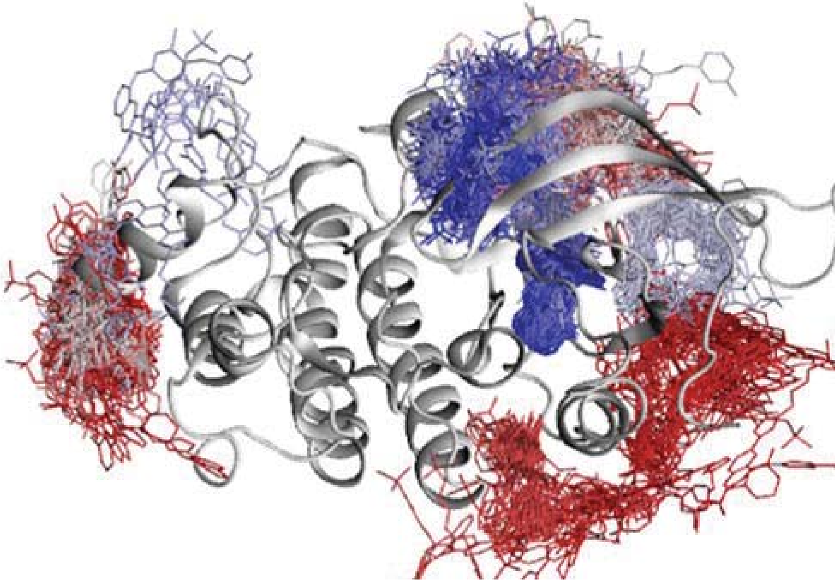
Internal coordinates





Source: Orengo, Christine A., and Janet M. Thornton. "Protein Families and their Evolution--A Structural Perspective." *Annual Review Biochemistry* 74 (2005): 867-900.

Comparing Structures



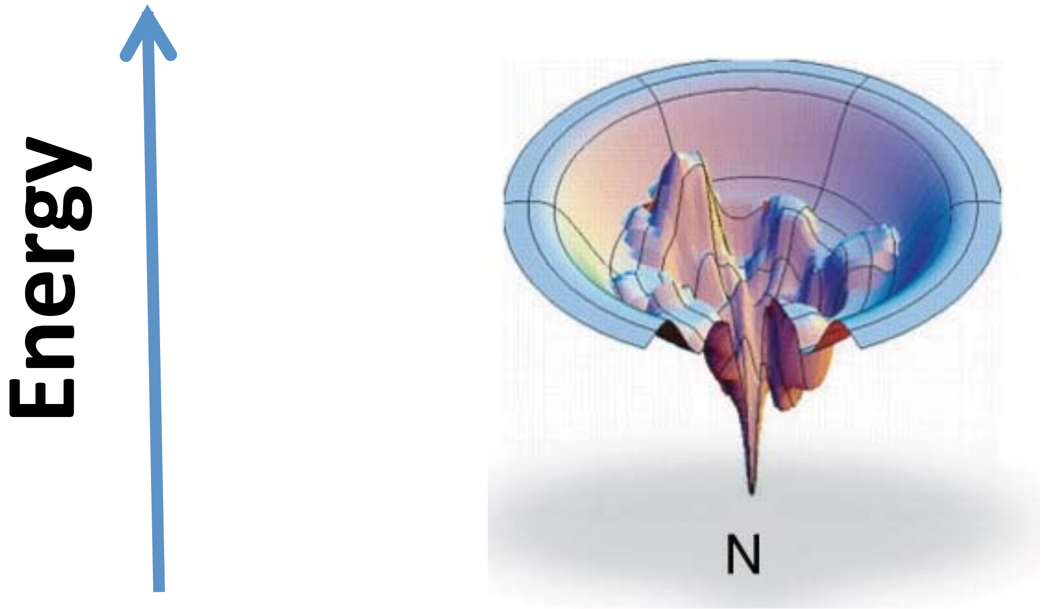
- Need to define corresponding atoms.
- Frequently only a subset of atoms:
  - main-chain
  - heavy atoms
- Minimize RMSD by rigid body transformations

$$RMSD(a,b)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(a_{ix}-b_{ix}\right)^2+\left(a_{iy}-b_{iy}\right)^2+\left(a_{iz}-b_{iz}\right)^2\right]}$$

QUESTIONS?

Protein Machines

Stable structure are energetic minima



Courtesy of Nature Publishing Group. Used with permission.
Source: Dill, Ken A., and Hue Sun Chan. "From Levinthal to Pathways to Funnels."
Nature Structural Biology 4, no. 1 (1997): 10-9.

$$F(\vec{x}) = -\nabla U(\vec{x})$$

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

Potential Energy of a Protein

Physicist

Statistician

Potential Energy of a Protein

Physicist

- Describe physical forces
- Equations may be approximate, but represent identifiable forces

CHARMM

Statistician

Potential Energy of a Protein

Physicist

- Describe physical forces
- Equations may be approximate, but represent identifiable forces

CHARMM

Statistician

- Describe observations
- No need to understand origin of statistical properties

Rosetta

Potential Energy of a Protein

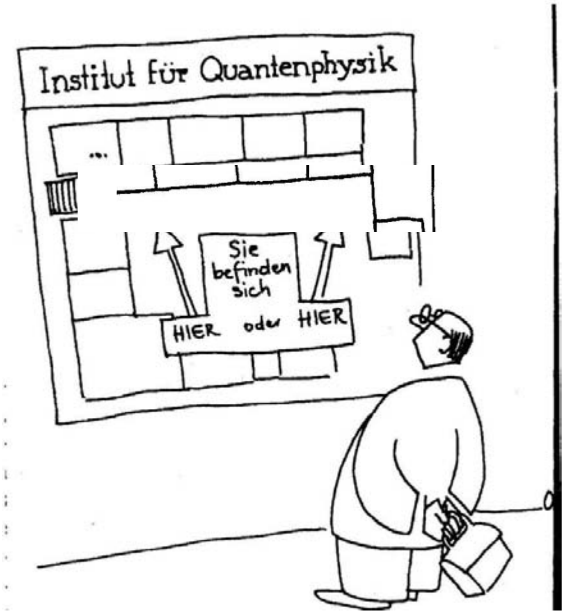
Physicist

- Describe physical forces
- Equations may be approximate, but represent identifiable forces

Statistician

- Describe observations
- No need to understand origin of statistical properties

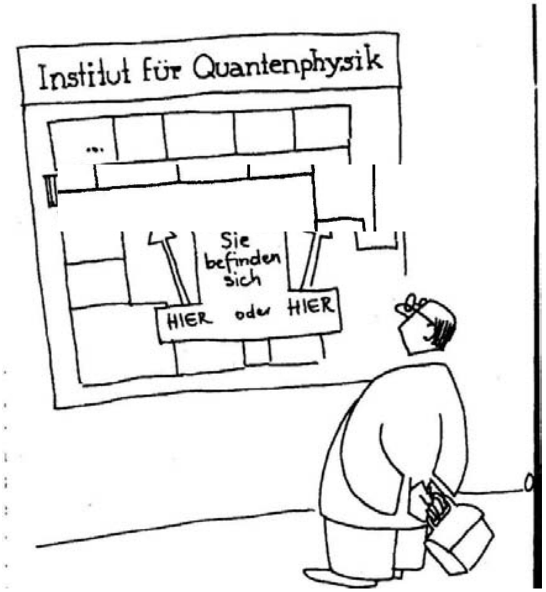
Rosetta



Potential Energy of a Protein

Physicist

- Describe physical forces
- Equations may be approximate, but represent identifiable forces



Statistician

- Describe observations
- No need to understand origin of statistical properties

“Data don’t make any sense, we will have to resort to statistics

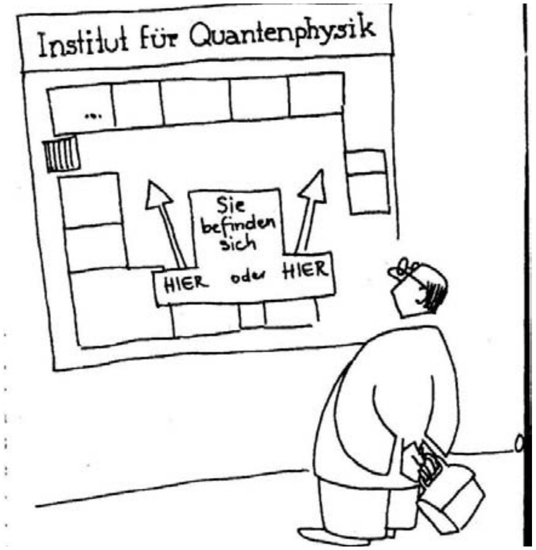
Courtesy of http://vadlo.com/. Used with permission.

CHARMM Energy Function

$$U_{CHARMM} = U_{bonded} + U_{non-bonded}$$

where U_{bonded} consists of the following terms,

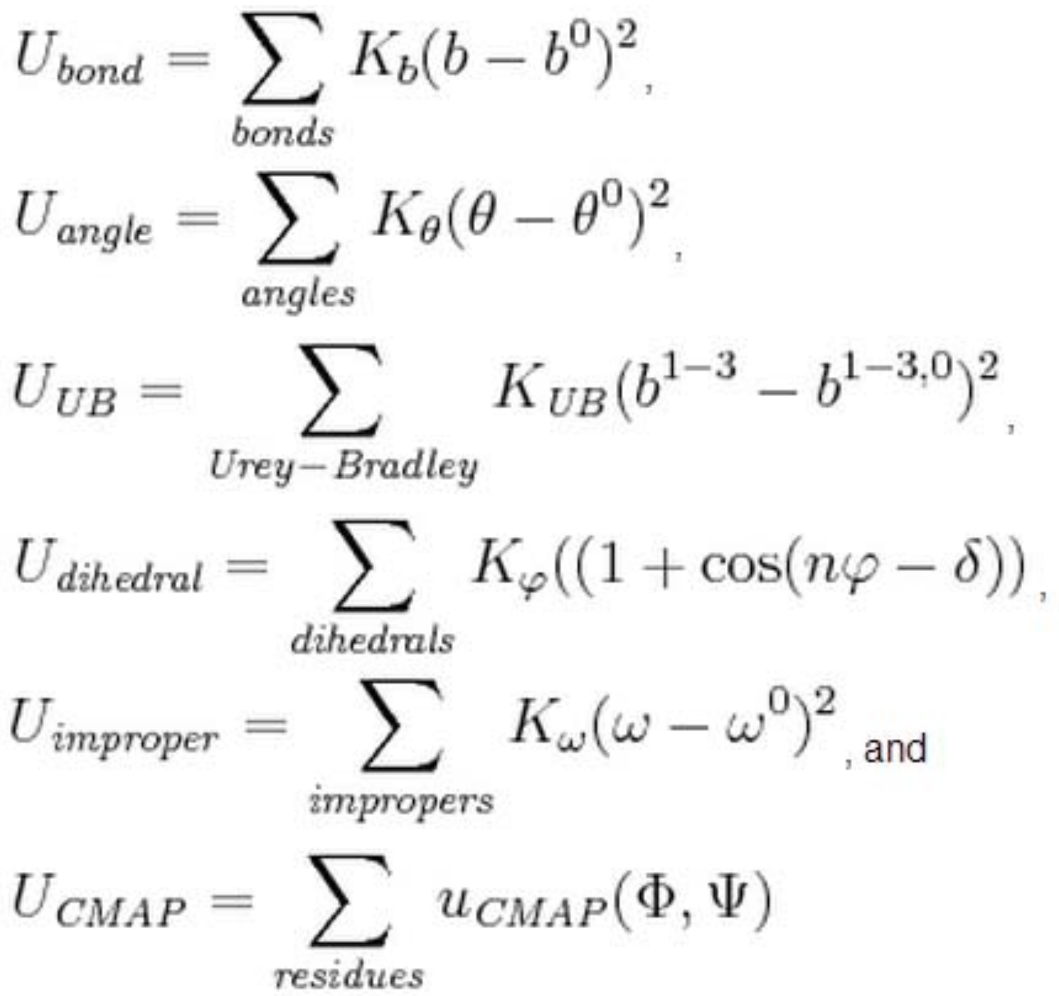
$$U_{bonded} = U_{bond} + U_{angle} + U_{UB} + U_{dihedral} + U_{improper} + U_{CMAP}$$



http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials

http://www.charmmtutorial.org/index.php/The_Energy_Function

CHARMM Energy Function U_{bonded}



$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2 ,$$

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta^0)^2 ,$$

$$U_{UB} = \sum_{Urey-Bradley} K_{UB}(b^{1-3} - b^{1-3,0})^2 ,$$

$$U_{dihedral} = \sum_{dihedrals} K_\varphi((1 + \cos(n\varphi - \delta)) ,$$

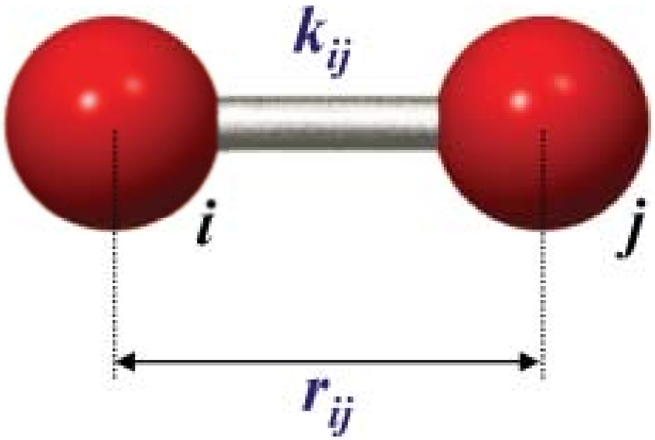
$$U_{improper} = \sum_{impropers} K_\omega(\omega - \omega^0)^2 \text{, and}$$

$$U_{CMAP} = \sum_{residues} u_{CMAP}(\Phi, \Psi)$$

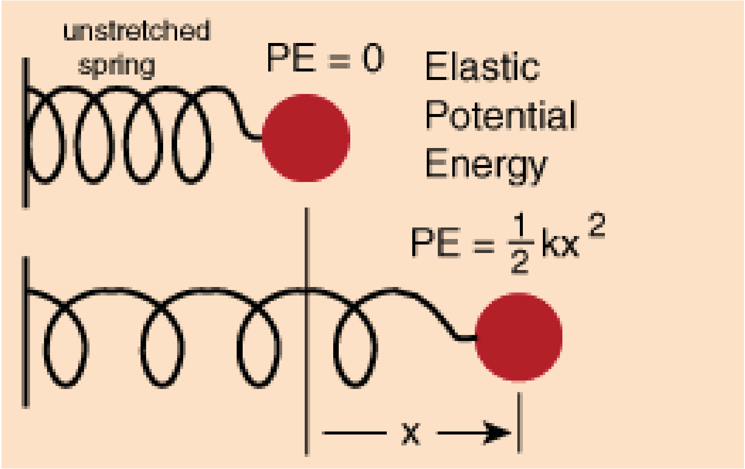
http://www.charmmtutorial.org/index.php/The_Energy_Function
http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials

CHARMM Energy Function U_{bonded}

$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2,$$



Harmonic forces maintain geometry



http://www.charmmtutorial.org/index.php/The_Energy_Function
http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials

CHARMM Energy Function U_{bonded}

$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2,$$

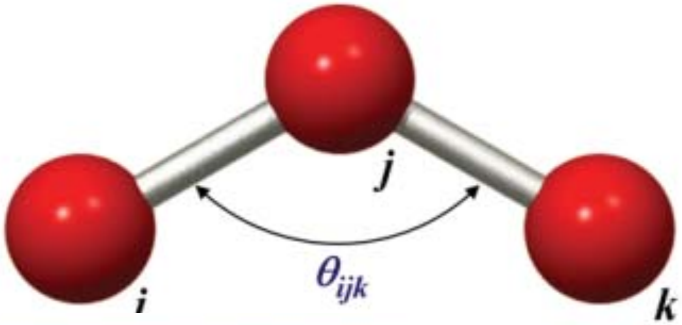
$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta^0)^2,$$

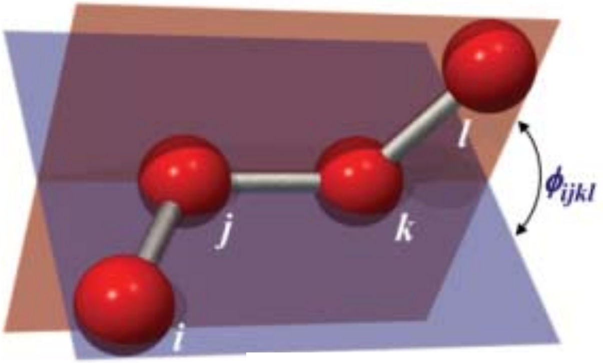
$$U_{UB} = \sum_{Urey-Bradley} K_{UB}(b^{1-3} - b^{1-3,0})^2,$$

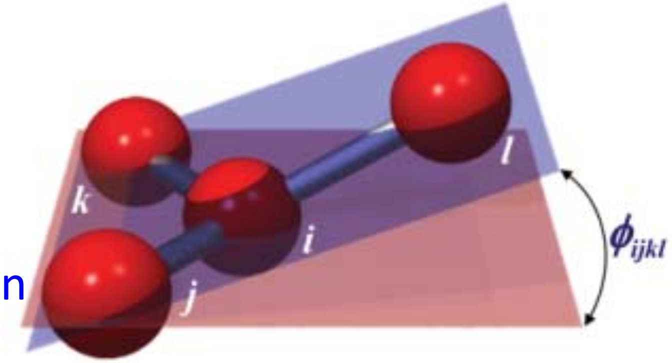
$$U_{dihedral} = \sum_{dihedrals} K_\varphi((1 + \cos(n\varphi - \delta)),$$

$$U_{improper} = \sum_{impropers} K_\omega(\omega - \omega^0)^2, \text{ and}$$

$$U_{CMAP} = \sum_{residues} u_{CMAP}(\Phi, \Psi)$$

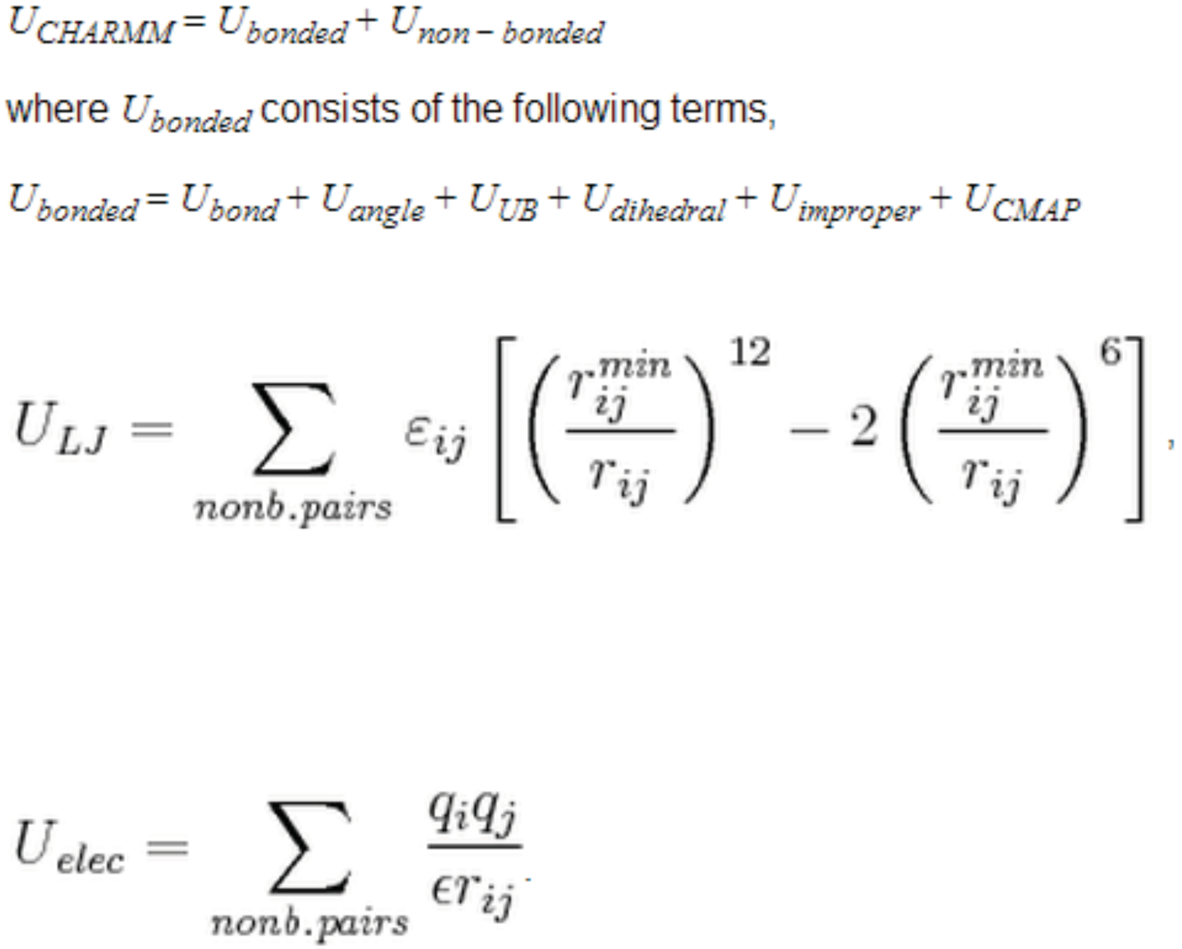






http://www.charmmtutorial.org/index.php/The_Energy_Function
http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials

CHARMM Energy Function $U_{\text{non-bonded}}$



$U_{CHARMM} = U_{bonded} + U_{non-bonded}$

where U_{bonded} consists of the following terms,

$U_{bonded} = U_{bond} + U_{angle} + U_{UB} + U_{dihedral} + U_{improper} + U_{CMAP}$

$$U_{LJ} = \sum_{nonb.pairs} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right],$$

$$U_{elec} = \sum_{nonb.pairs} \frac{q_i q_j}{\epsilon r_{ij}}.$$

Non-bonded terms:
Lennard Jones

and

Electrostatics

CHARMM Energy Function $U_{non\text{-}bonded}$

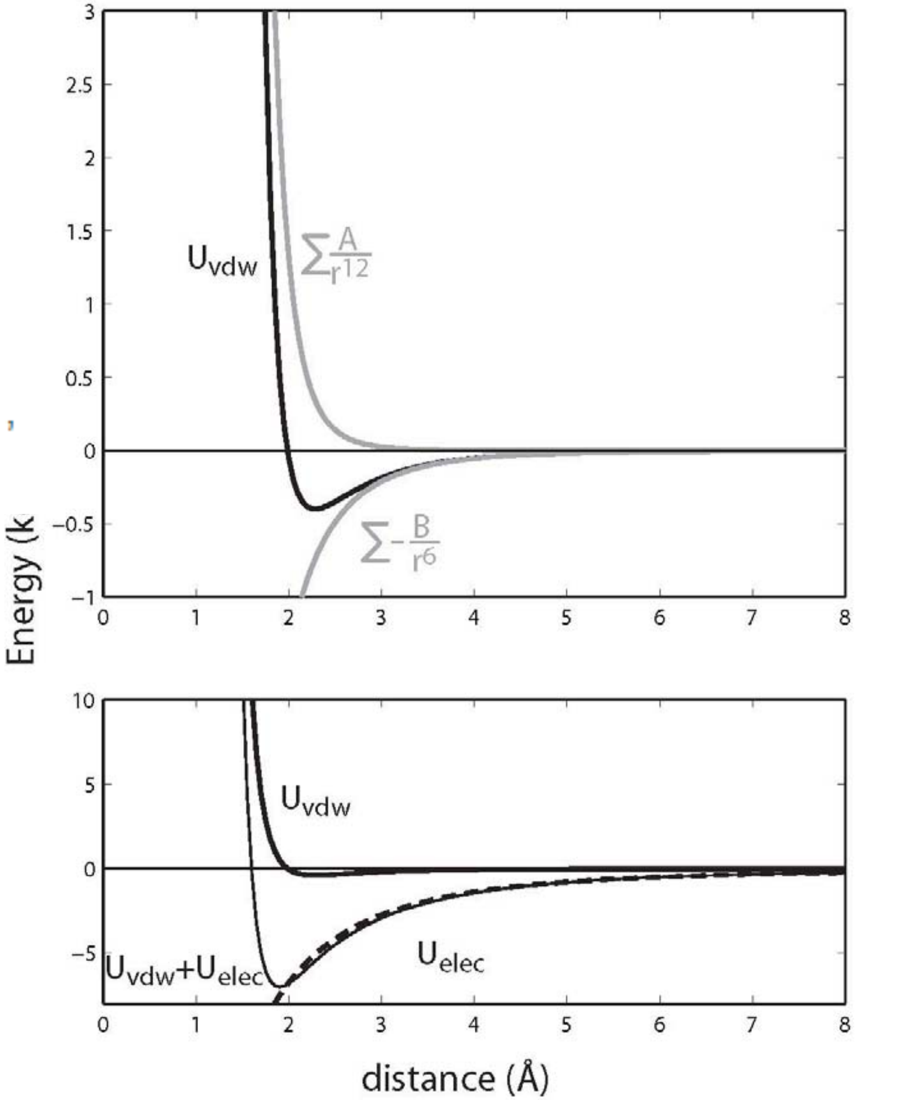
$$U_{CHARMM} = U_{bonded} + U_{non-bonded}$$

where U_{bonded} consists of the following terms,

$$U_{bonded} = U_{bond} + U_{angle} + U_{UB} + U_{dihedral} + U_{improper} + U_{CMAP}$$

$$U_{LJ} = \sum_{nonb.pairs} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right],$$

$$U_{elec} = \sum_{nonb.pairs} \frac{q_i q_j}{\epsilon r_{ij}}.$$

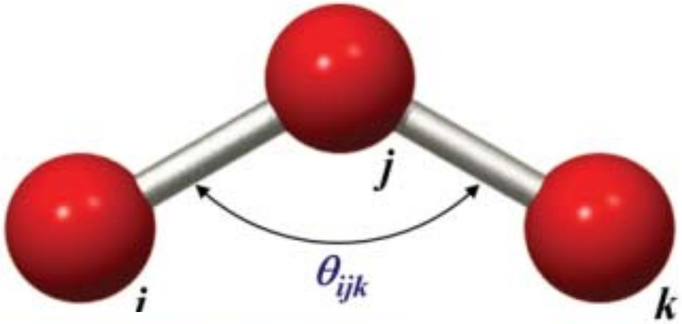


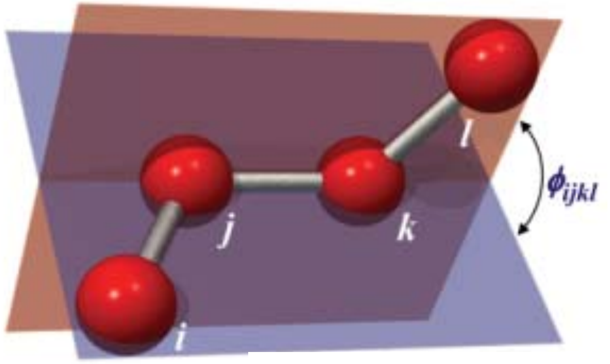
QUESTIONS?

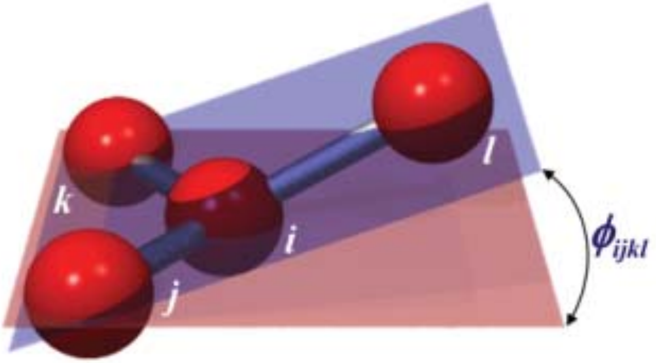
Rosetta Energy Function

Keep geometry fixed!

$$\cancel{\begin{aligned}
U_{bond} &= \sum_{bonds} K_b(b-b^0)^2, \\
U_{angle} &= \sum_{angles} K_\theta(\theta-\theta^0)^2, \\
U_{UB} &= \sum_{Urey-Bradley} K_{UB}(b^{1-3}-b^{1-3,0})^2, \\
U_{dihedral} &= \sum_{dihedrals} K_\varphi(1+\cos(n\varphi-\delta)), \\
U_{improper} &= \sum_{impropers} K_\omega(\omega-\omega^0)^2 \text{, and} \\
U_{CMAP} &= \sum_{residues} u_{CMAP}(\Phi,\Psi)
\end{aligned}}$$

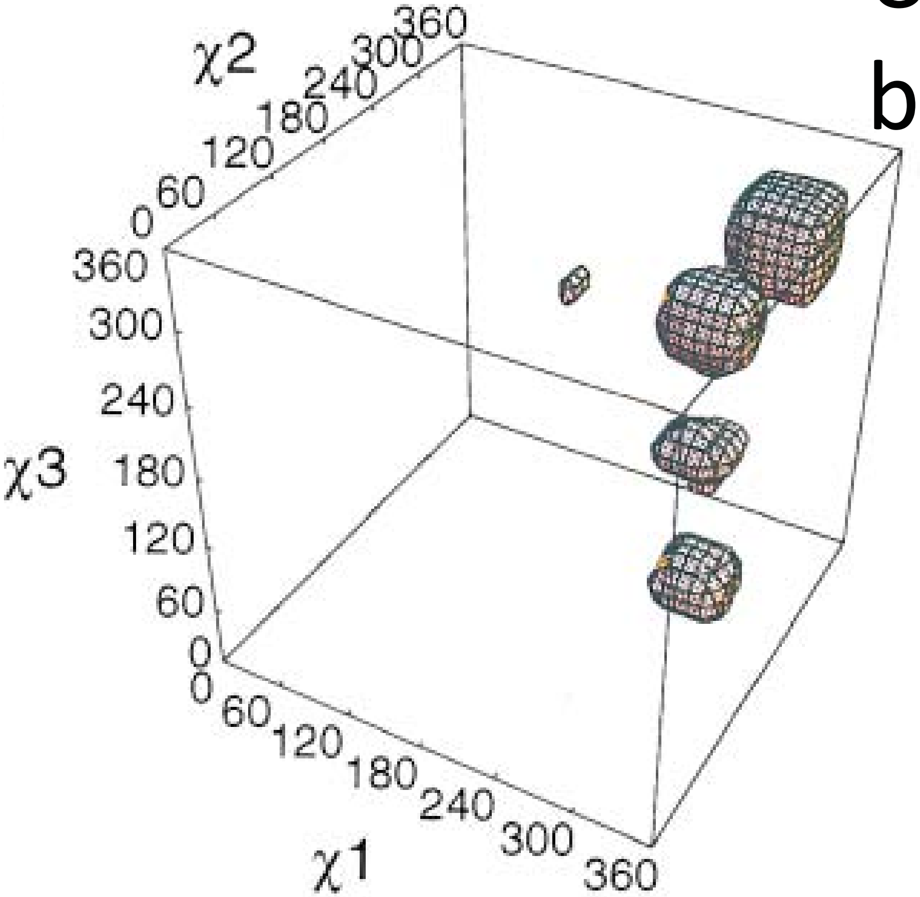


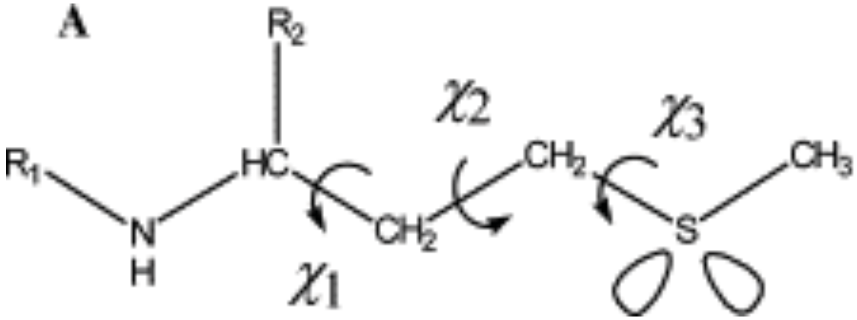




Rosetta Energy Function

Rotamers:
Use discrete angles when bonds rotate

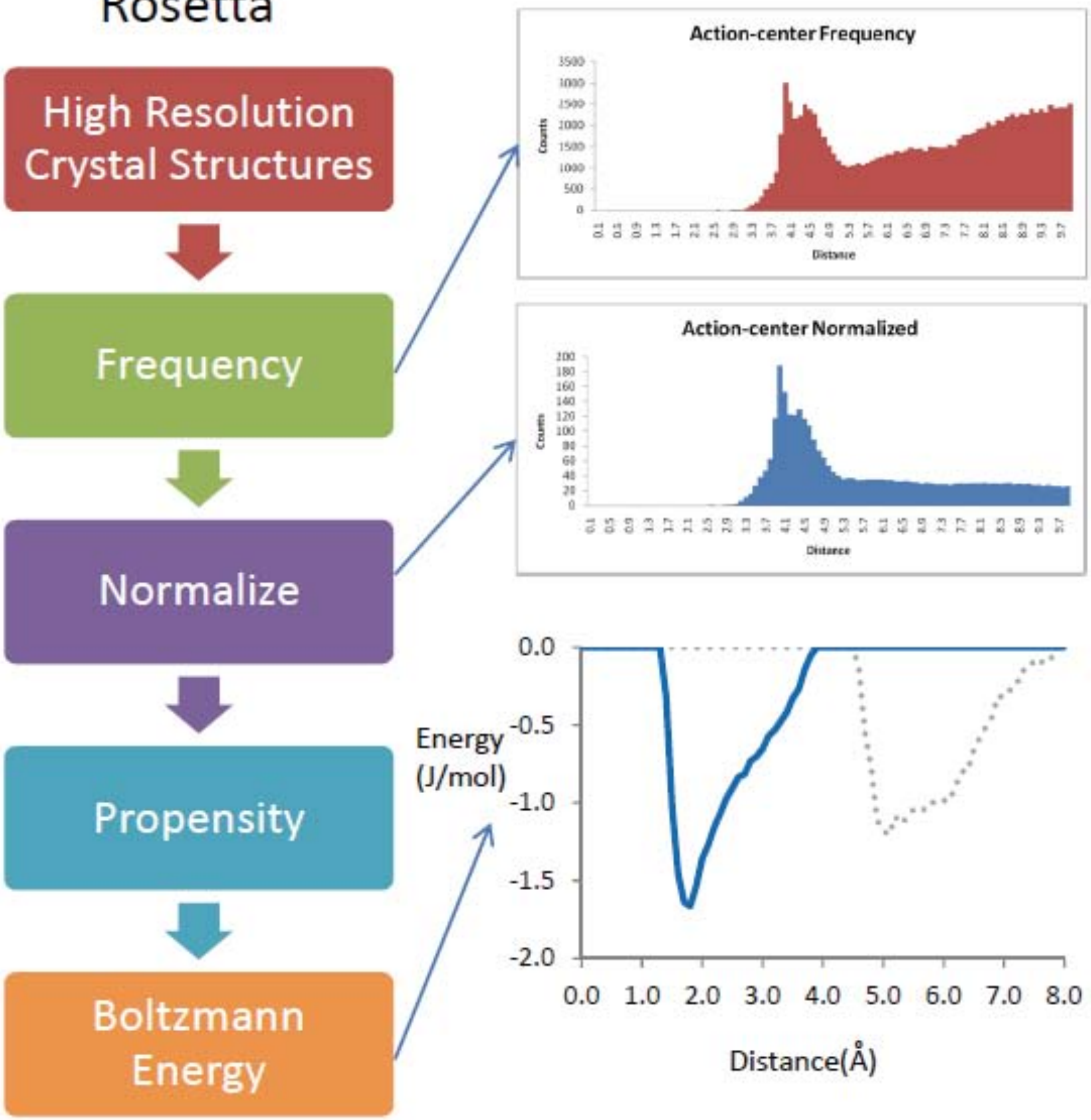




Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Kuszewski, John, Angela M. Gronenborn, et al. "Improvements and Extensions in the Conformational Database Potential for the Refinement of NMR and X-ray Structures of Proteins and Nucleic Acids." *Journal of Magnetic Resonance* 125, no. 1 (1997): 171-7.

Knowledge Based

Rosetta



High Resolution Crystal Structures

Frequency

Normalize

Propensity

Boltzmann Energy

Frequency of states

$$g_{ij}(r) = \rho_{ij}(r)/\rho^*_{ij}(r)$$

Empirical potential energy

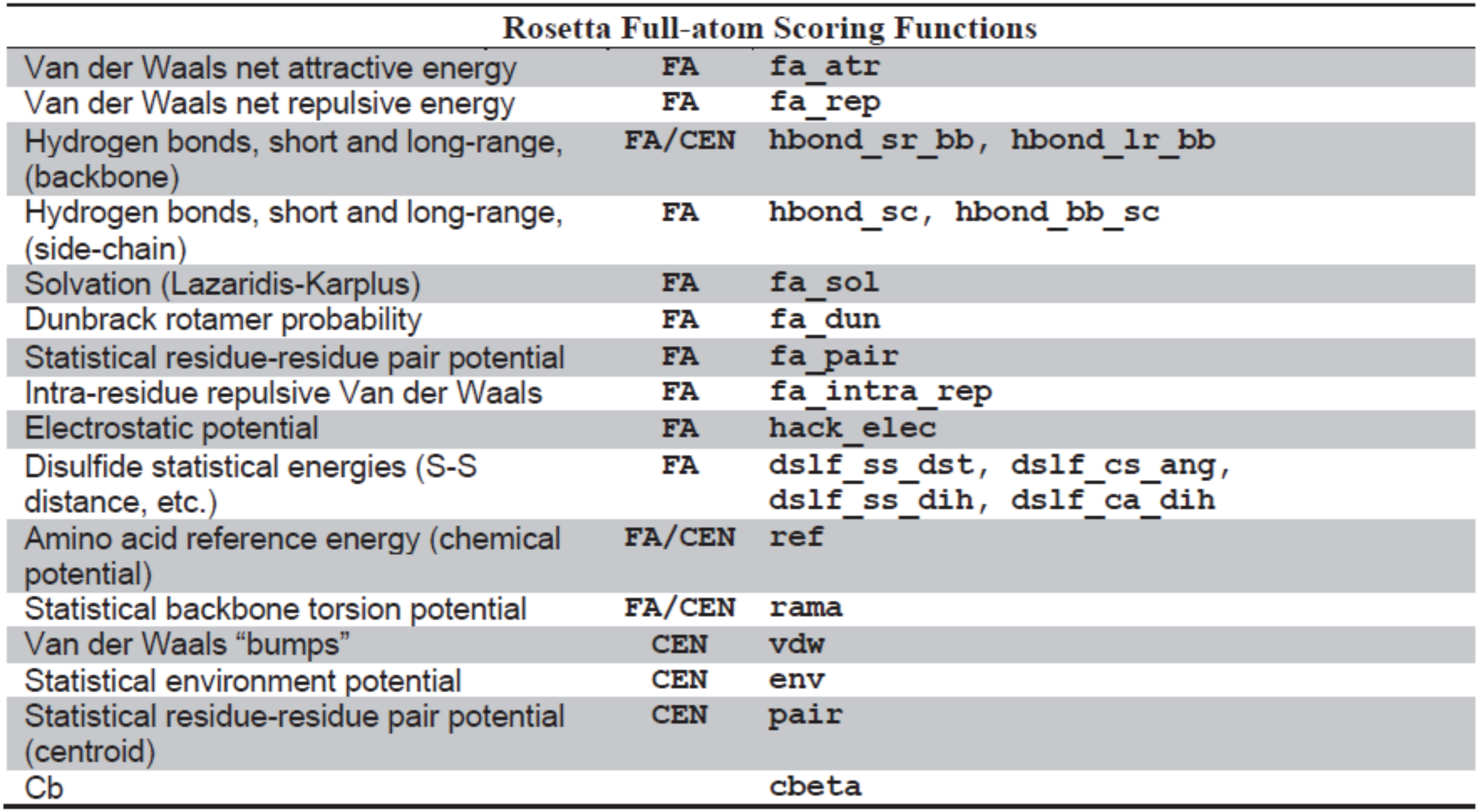
$$u_{ij}(r) = -k_B T \ln[g_{ij}(r)]$$

*Energy = w₁*term₁ + w₂*term₂ + …*

Courtesy of Steven Combs (PDF). Used with permission.

3.3 Scoring components

The most common score function components are:



Rosetta Full-atom Scoring Functions

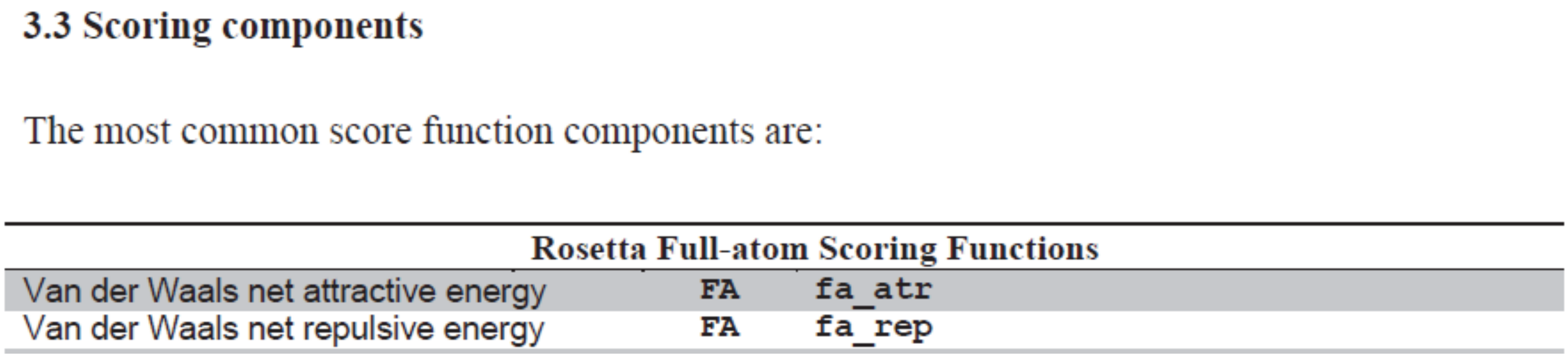
Van der Waals net attractive energy	**FA**	`fa_atr`
Van der Waals net repulsive energy	**FA**	`fa_rep`
Hydrogen bonds, short and long-range, (backbone)	**FA/CEN**	`hbond_sr_bb, hbond_lr_bb`
Hydrogen bonds, short and long-range, (side-chain)	**FA**	`hbond_sc, hbond_bb_sc`
Solvation (Lazaridis-Karplus)	**FA**	`fa_sol`
Dunbrack rotamer probability	**FA**	`fa_dun`
Statistical residue-residue pair potential	**FA**	`fa_pair`
Intra-residue repulsive Van der Waals	**FA**	`fa_intra_rep`
Electrostatic potential	**FA**	`hack_elec`

Disulfide statistical energies (S-S distance, etc.)	**FA**	`dslf_ss_dst, dslf_cs_ang,`
		`dslf_ss_dih, dslf_ca_dih`

Amino acid reference energy (chemical potential)	**FA/CEN**	`ref`
Statistical backbone torsion potential	**FA/CEN**	`rama`
Van der Waals "bumps"	**CEN**	`vdw`
Statistical environment potential	**CEN**	`env`
Statistical residue-residue pair potential (centroid)	**CEN**	`pair`
Cb		`cbeta`

Courtesy of Jeffrey J Gray (PDF). Used with permission.

Note that a number of scoring components are compatible with both full-atom and centroid mode.



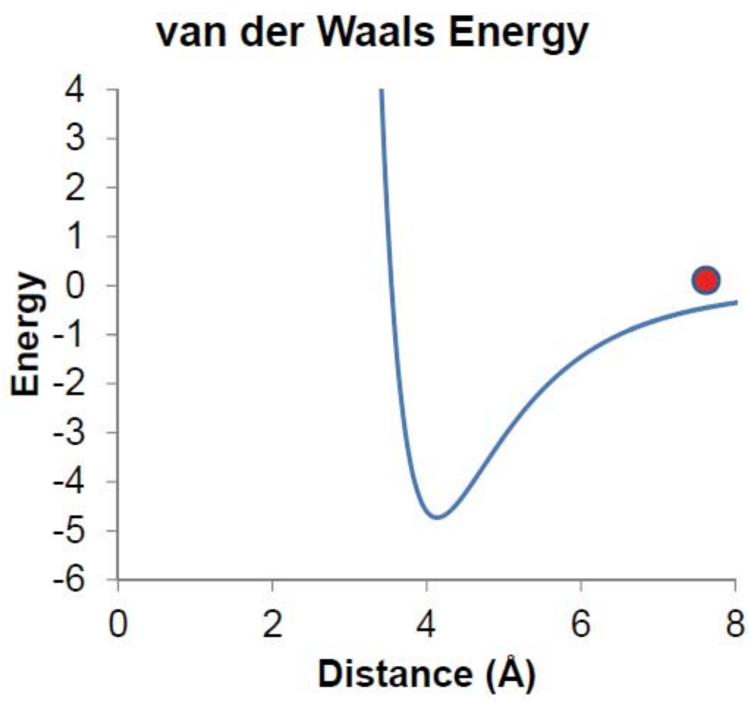
3.3 Scoring components

The most common score function components are:

Rosetta Full-atom Scoring Functions		
Van der Waals net attractive energy	**FA**	**fa_atr**
Van der Waals net repulsive energy	**FA**	**fa_rep**

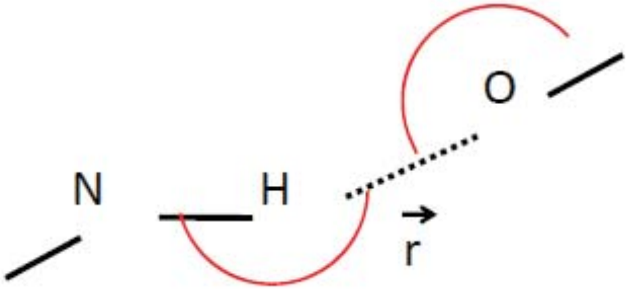
Courtesy of Jeffrey J Gray (PDF) . Used with permission.

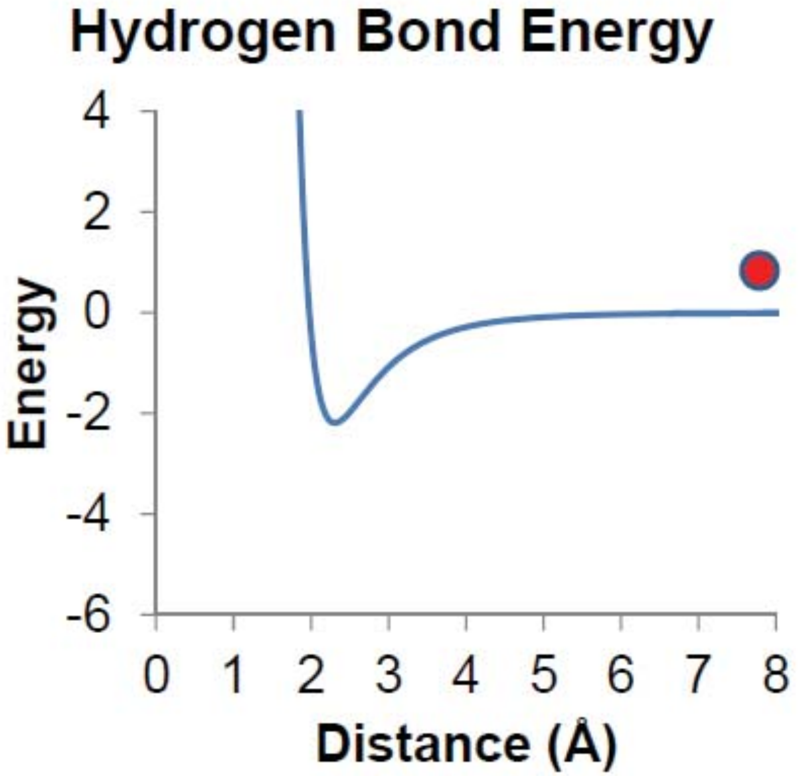
very similar to physicist view



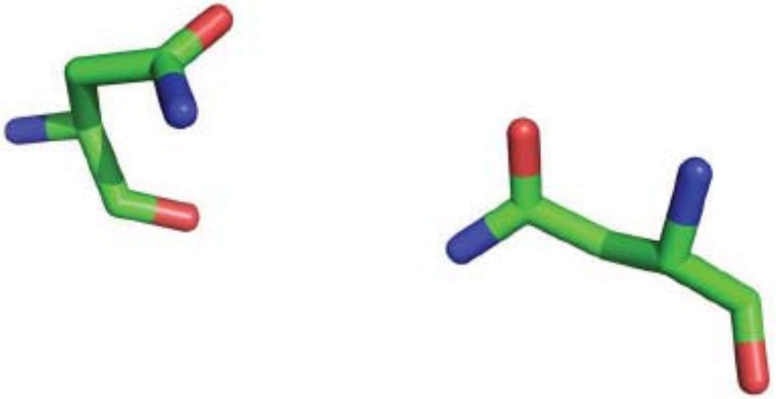
Courtesy of Steven Combs (PDF). Used with permission.

- `Hbond_lr_bb / hbond_sr_bb / hbond_bb_sc / hbond_sc`
- Geometry dependent
  - 2 angles, 1 distance
- Lives in: `src/core/scoring/hbonds/HbondEnergy.cc`



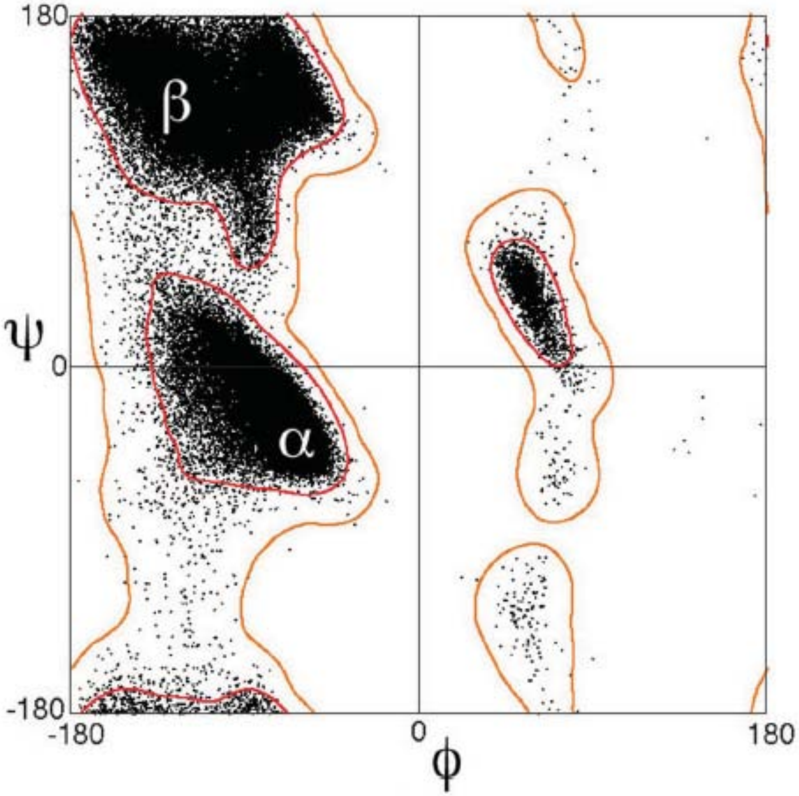


Courtesy of Steven Combs (PDF). Used with permission.

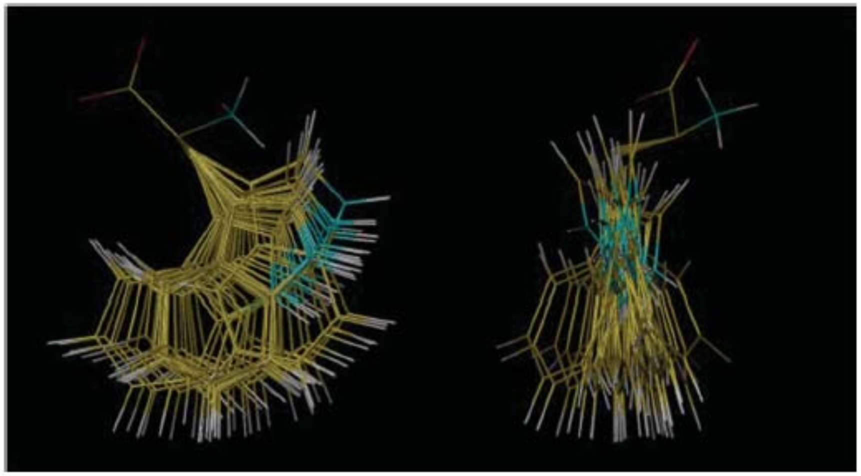


Animation by:
Kristian Kaufmann

Prefer common rotations

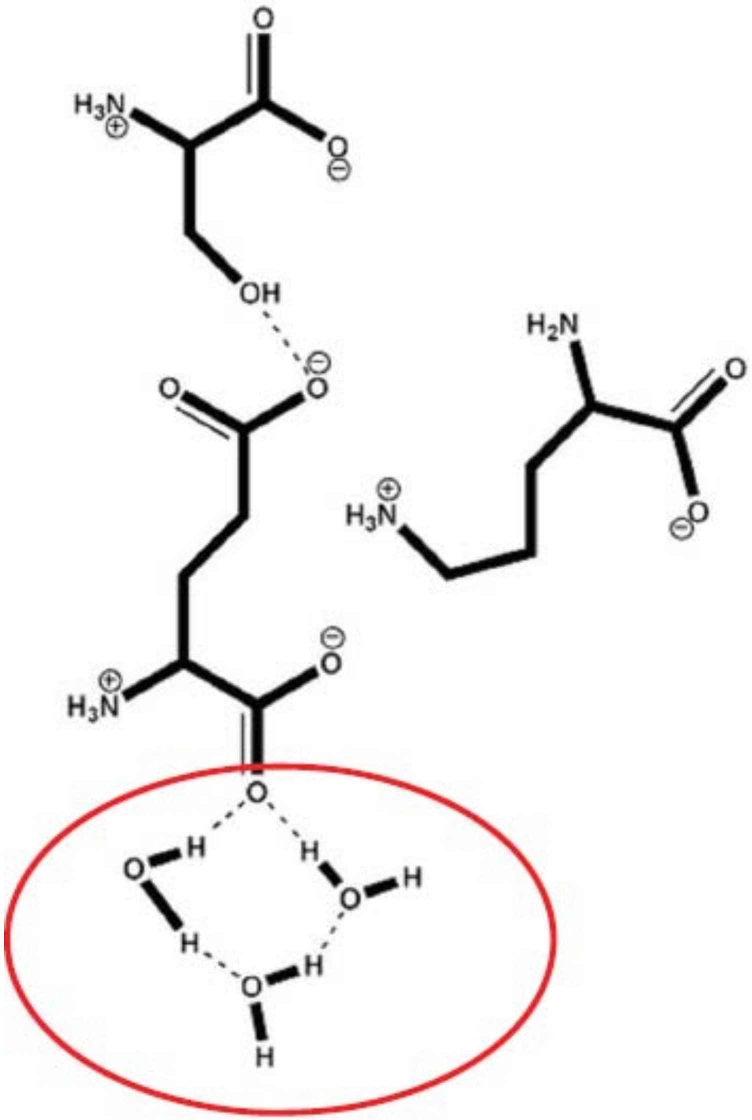


Courtesy of Jane S. Richardson. License: CC-BY.



Source: Mendes, Joaquim, António M. Baptista, et al. "Improved Modeling of Side-chains in Proteins with Rotamer-based Methods: A Flexible Rotamer Model." *Proteins: Structure, Function, and Bioinformatics* 37, no. 4 (1999): 530-43.

Solvation is very hard for the physicist



Hydration Shell

Solvation is very hard for the physicist, easy for the statistician

Empirical solution

$$\Delta G_i^{solv} = \Delta G_i^{\mathrm{Ref}} - \sum_{j \neq i} f_i(r_{ij}) V_j$$

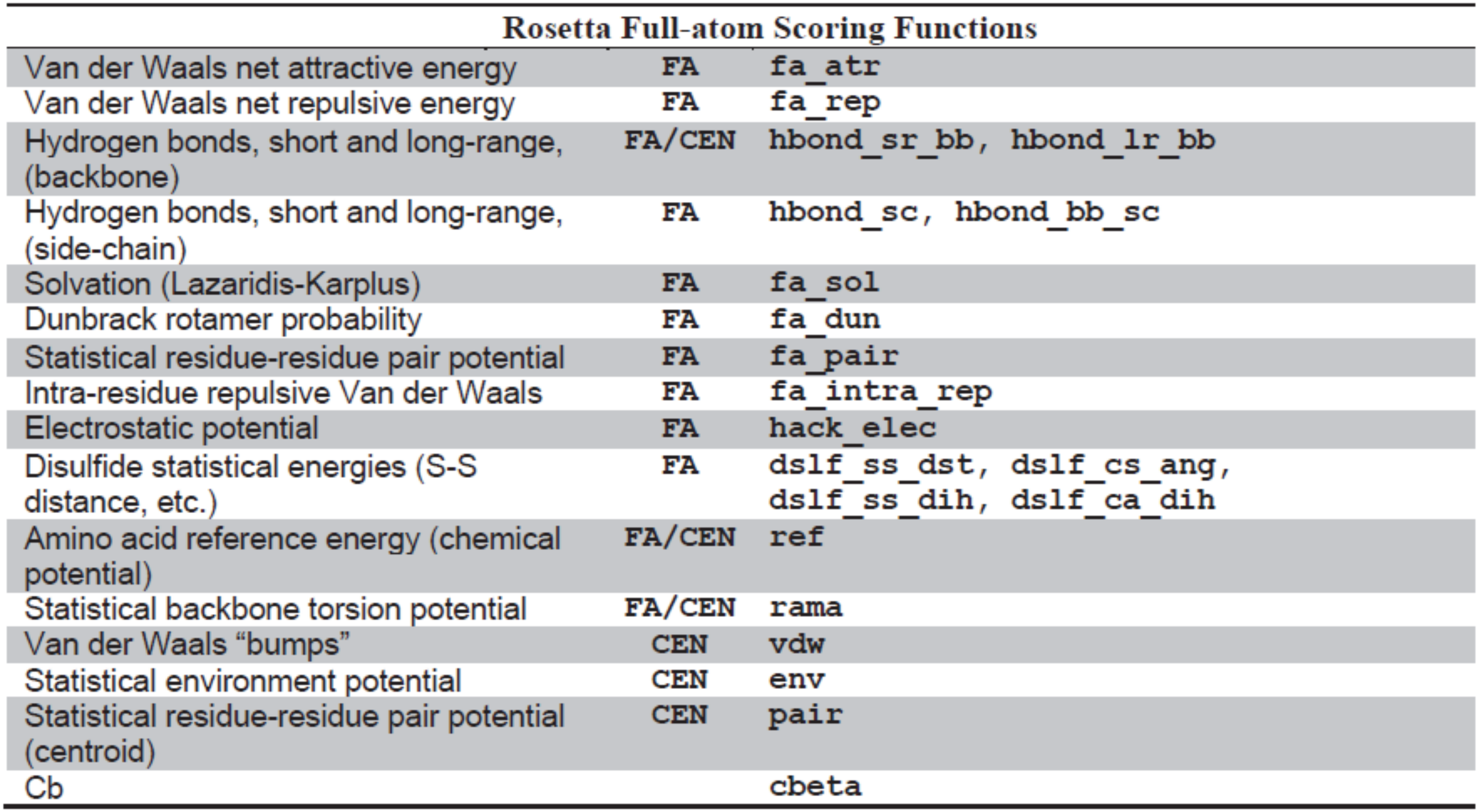
Experimentally determined solvation of group when fully solvent exposed. (From transfer experiments)

Distance-dependent function for interaction of groups i,j

Volume of neighboring group j

3.3 Scoring components

The most common score function components are:



Rosetta Full-atom Scoring Functions

Van der Waals net attractive energy	**FA**	**`fa_atr`**
Van der Waals net repulsive energy	**FA**	**`fa_rep`**
Hydrogen bonds, short and long-range, (backbone)	**FA/CEN**	**`hbond_sr_bb, hbond_lr_bb`**
Hydrogen bonds, short and long-range, (side-chain)	**FA**	**`hbond_sc, hbond_bb_sc`**
Solvation (Lazaridis-Karplus)	**FA**	**`fa_sol`**
Dunbrack rotamer probability	**FA**	**`fa_dun`**
Statistical residue-residue pair potential	**FA**	**`fa_pair`**
Intra-residue repulsive Van der Waals	**FA**	**`fa_intra_rep`**
Electrostatic potential	**FA**	**`hack_elec`**

Disulfide statistical energies (S-S distance, etc.)	**FA**	**`dslf_ss_dst, dslf_cs_ang,`**
		`dslf_ss_dih, dslf_ca_dih`

Amino acid reference energy (chemical potential)	**FA/CEN**	**`ref`**
Statistical backbone torsion potential	**FA/CEN**	**`rama`**
Van der Waals "bumps"	**CEN**	**`vdw`**
Statistical environment potential	**CEN**	**`env`**
Statistical residue-residue pair potential (centroid)	**CEN**	**`pair`**
Cb		**`cbeta`**

Courtesy of Jeffrey J Gray (PDF). Used with permission.

Note that a number of scoring components are compatible with both full-atom and centroid mode.

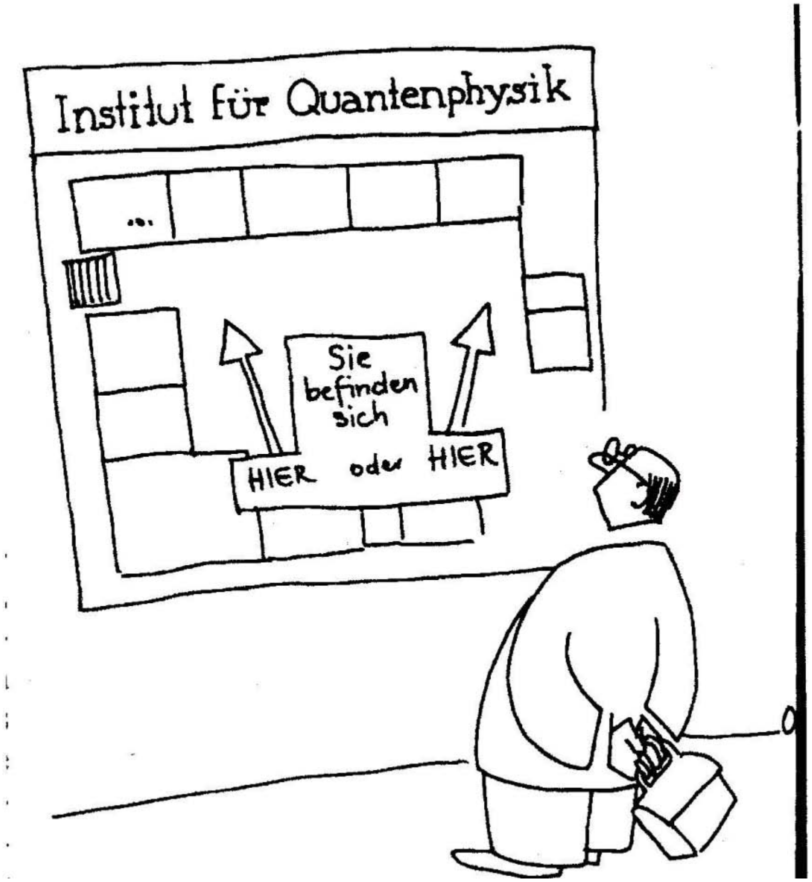
Summary

- Protein structure influences all biology
- Experimental techniques give constraints, not structures
- Computational methods needed to interpret constraints
- Two main approaches: physical and statistical

What were the key simplifications of the statistical approach?

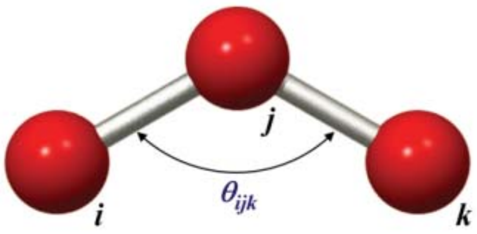
"Data don't make any sense, we will have to resort to statistics."

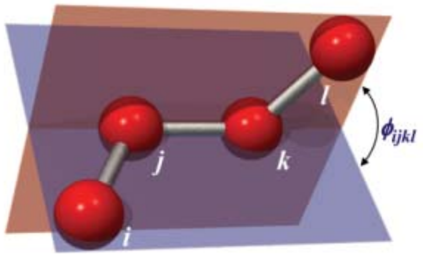
Courtesy of http://vadlo.com/. Used with permission.



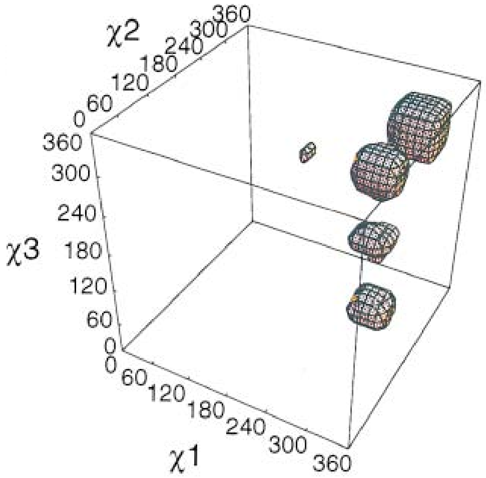
What were the key simplifications of the statistical approach?

- Fixed geometry





- Discrete rotamers



- Statistical potential

Frequency of states

$$g_{ij}(r) = \rho_{ij}(r)/\rho^{*}_{ij}(r)$$

Empirical potential energy

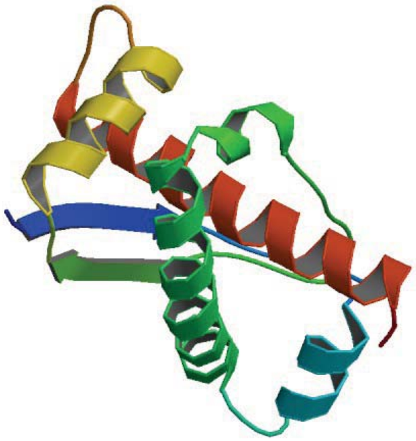
$$u_{ij}(r) = -k_{\mathrm{B}}T\ln[g_{ij}(r)]$$

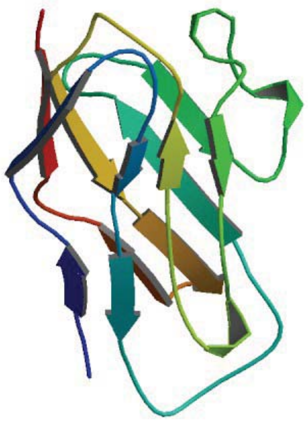
Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Kuszewski, John, Angela M. Gronenborn, et al. "Improvements and Extensions in the Conformational Database Potential for the Refinement of NMR and X-ray Structures of Proteinsand NucleicAcids." *Journal of Magnetic Resonance* 125, no. 1 (1997): 171-7.

A thought experiment:
Which structure matches a sequence?

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

?

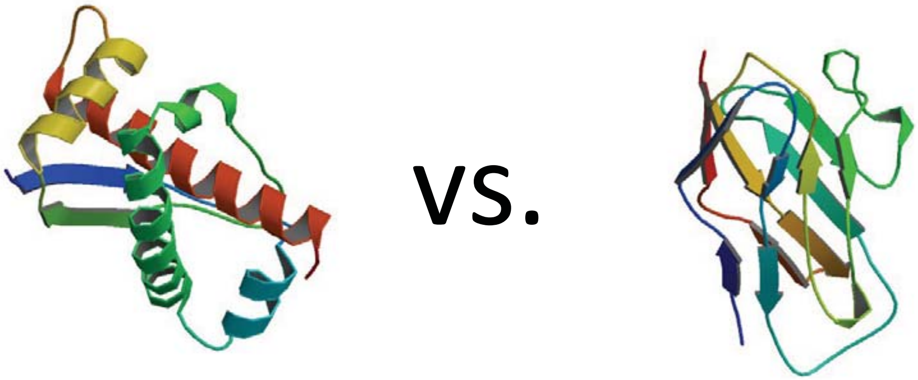




http://www.rcsb.org/pdb/images/2rh3_bio_r_500.jpg

Courtesy of RCSB Protein Data Bank. Used with permission.

http://www.rcsb.org/pdb/images/1qfp_bio_r_500.jpg



vs.

Courtesy of RCSB Protein Data Bank. Used with permission.

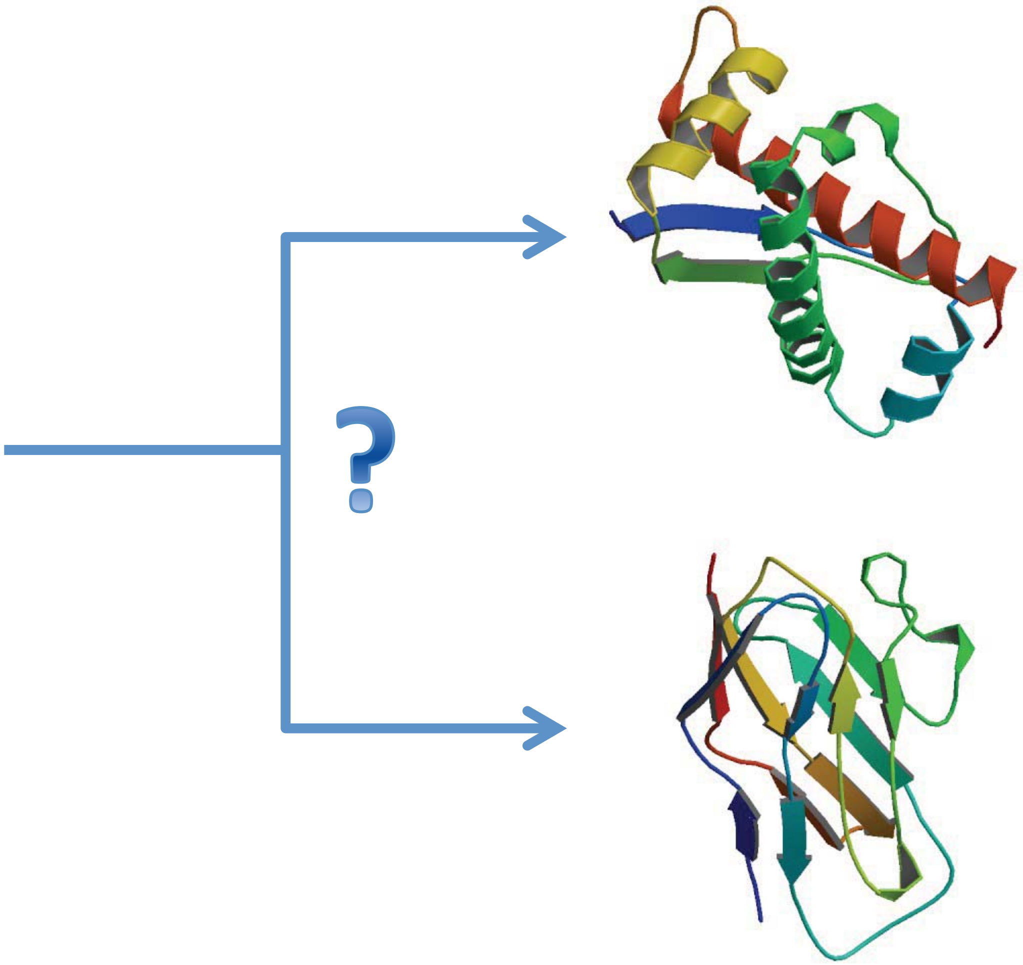
- How could you use energy functions to distinguish?
  - Let's assume one of the structures is the correct one.
  - Which should have the lower potential energy?
  - What do you think happens in practice?

Courtesy of RCSB Protein Data Bank. Used with permission.

- If one of the structures is the correct one:
  - Need to determine side chain conformations before calculating potential
- If better structure is only approximate:
  - Need to refine backbone and side chains first.

Threading (fold recognition)

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

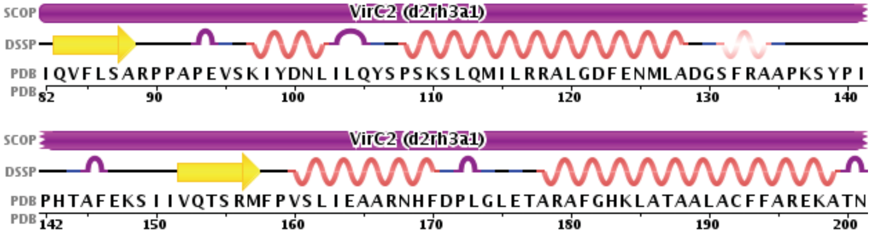


Courtesy of RCSB Protein Data Bank.
Used with permission.

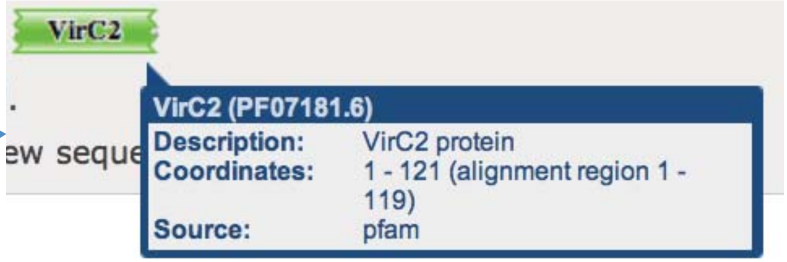
Other prediction problems

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

secondary structure


SCOP
VirC2 (d2rh3a1)
DSSP
PDB I Q V F L S A R P P A P E V S K I Y D N L I L Q Y S P S K S L Q M I L R R A L G D F E N M L A D G S F R A A P K S Y P I
PDB 82 90 100 110 120 130 140
SCOP
VirC2 (d2rh3a1)
DSSP
PDB P H T A F E K S I I V Q T S R M F P V S L I E A A R N H F D P L G L E T A R A F G H K L A T A A L A C F F A R E K A T N
PDB 142 150 160 170 180 190 200

domain structure



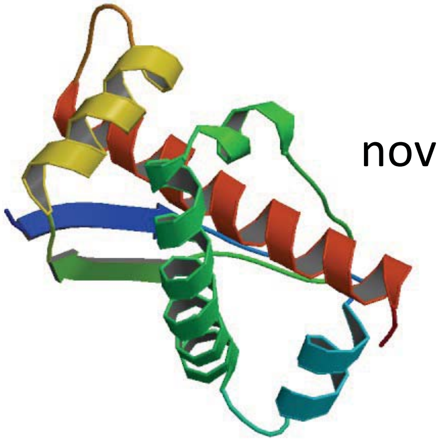
VirC2 (PF07181.6)

Description: VirC2 protein

Coordinates: 1 - 121 (alignment region 1 - 119)

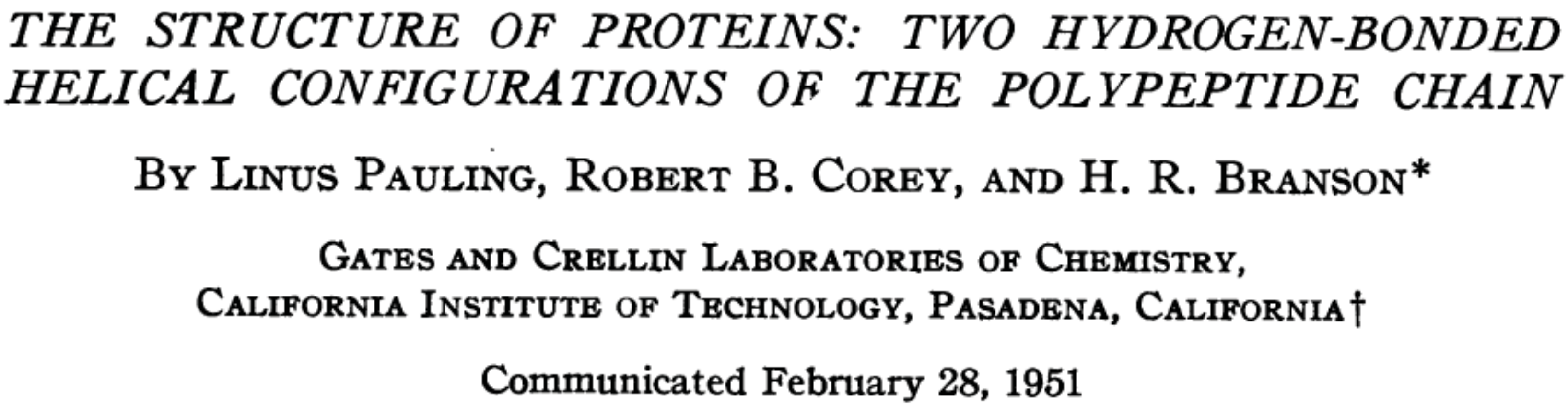
Source: pfam

novel 3D structure



Courtesy of RCSB Protein Data Bank.
Used with permission.

Some history



THE STRUCTURE OF PROTEINS: TWO HYDROGEN-BONDED HELICAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN

BY LINUS PAULING, ROBERT B. COREY, AND H. R. BRANSON*

GATES AND CRELLIN LABORATORIES OF CHEMISTRY, CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA†

Communicated February 28, 1951

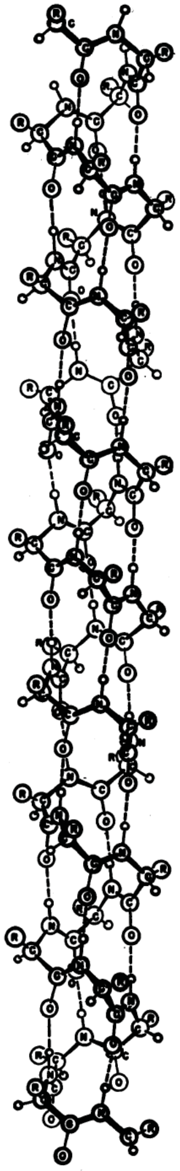
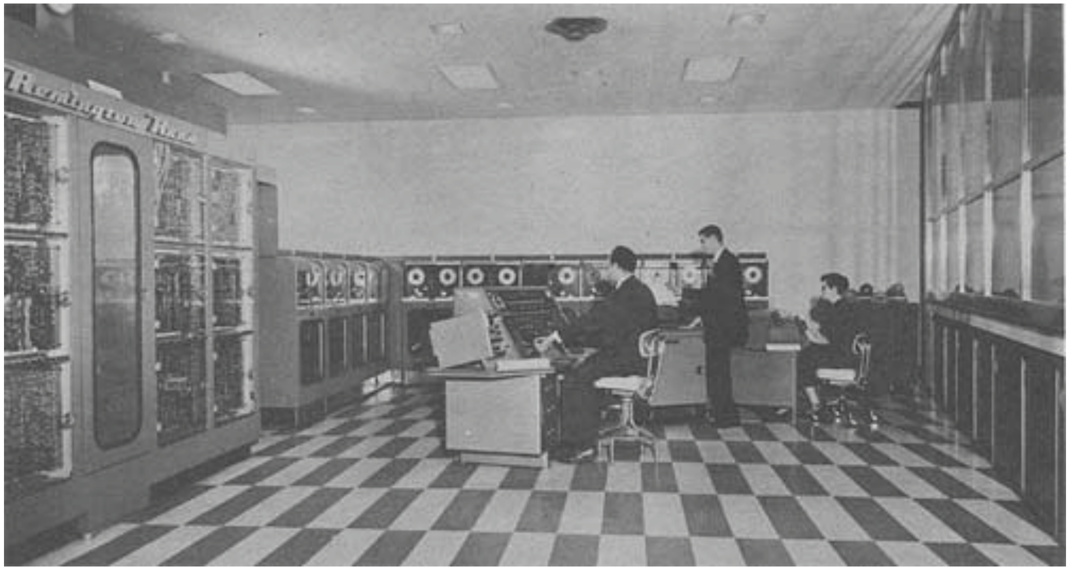


FIGURE 2

The helix with 3.7 residues per turn.

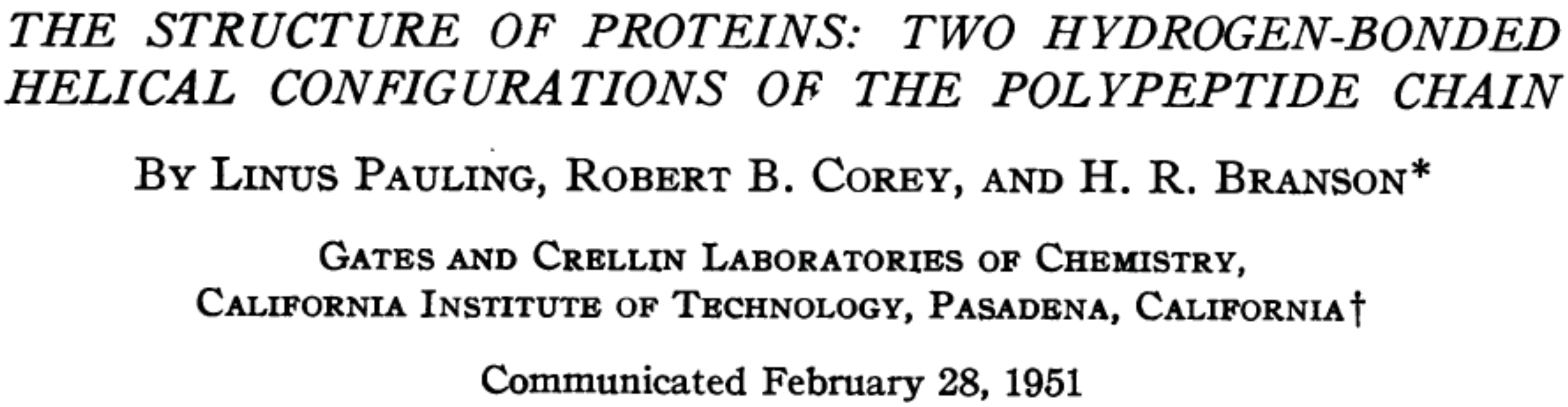


Courtesy of U.S. Department of the Army Ballistic Research Report. In the public domain.

UNIVAC 1 released in 1951

Source: Pauling, Linus, Robert B. Corey, and Herman R. Branson. "The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain." *Proceedings of the National Academy of Sciences* 37, no. 4 (1951): 205-211.

Some history



THE STRUCTURE OF PROTEINS: TWO HYDROGEN-BONDED HELICAL CONFIGURATIONS OF THE POLYPEPTIDE CHAIN

BY LINUS PAULING, ROBERT B. COREY, AND H. R. BRANSON*

GATES AND CRELLIN LABORATORIES OF CHEMISTRY,
CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA†

Communicated February 28, 1951

- Paper models!
- Key insight while lying in bed, sick
- Preceded by lots of hard work collecting experimental data
- Planar peptide bonds
- Maximize hydrogen bonds

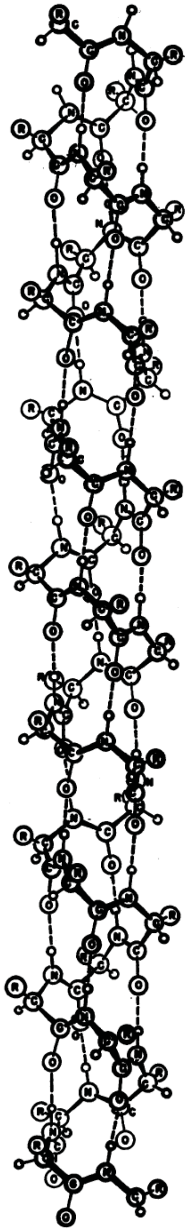


FIGURE 2
The helix with 3.7 residues per turn.

Source: Pauling, Linus, Robert B. Corey, and Herman R. Branson. "The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain." *Proceedings of the National Academy of Sciences* 37, no. 4 (1951): 205-211.

J. Mol. Biol. (1963) **7**, 95–99

Stereochemistry of Polypeptide Chain Configurations

Department of Physics
University of Madras
Madras 25, India

G. N. Ramachandran
C. Ramakrishnan
V. Sasisekharan

Received 27 December 1962

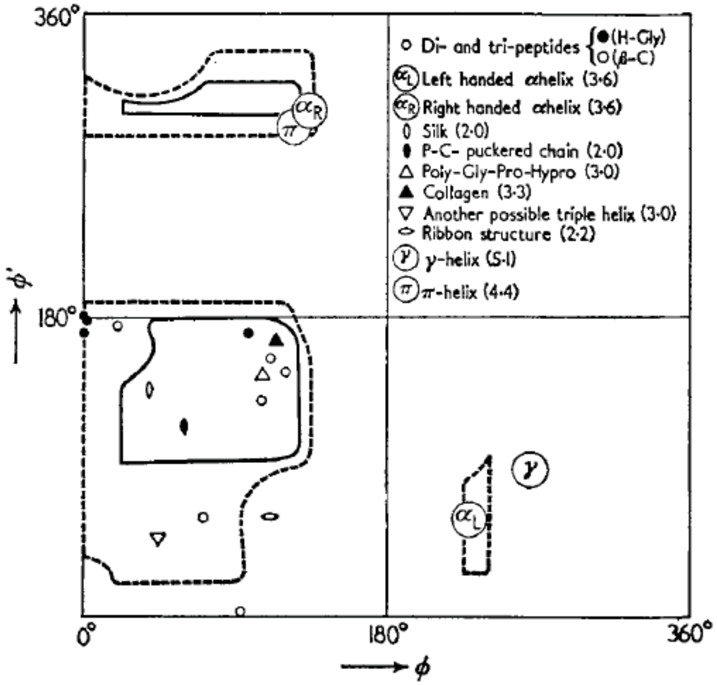
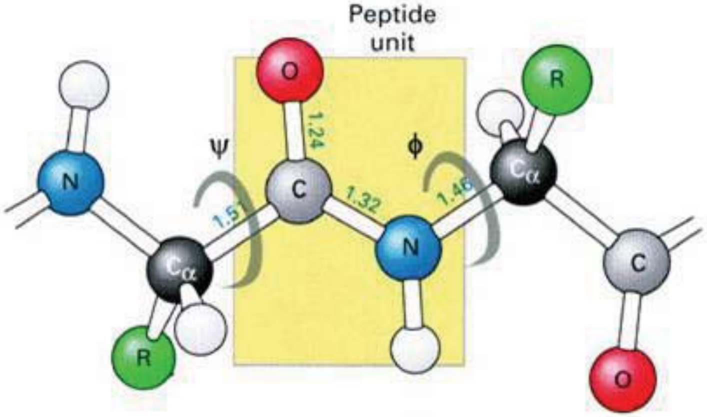
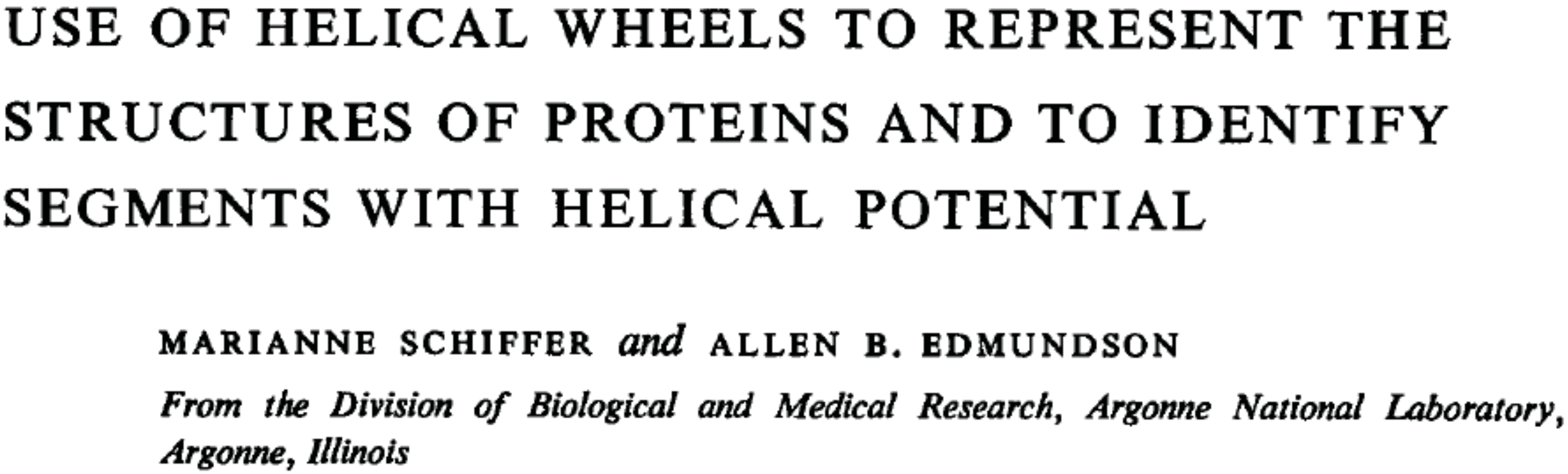


Fig. 3. Contours of constant n (———) and constant h (– – – – – –) corresponding to the angle N—αC—C′ = 110°. The boundaries of the fully allowed and outer limit regions are also shown.

Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Ramachandran, G. N., C. T. Ramakrishnan, et al. "Stereochemistry of Polypeptide Chain Configurations." *Journal of Molecular Biology* 7, no. 1 (1963): 95-9.



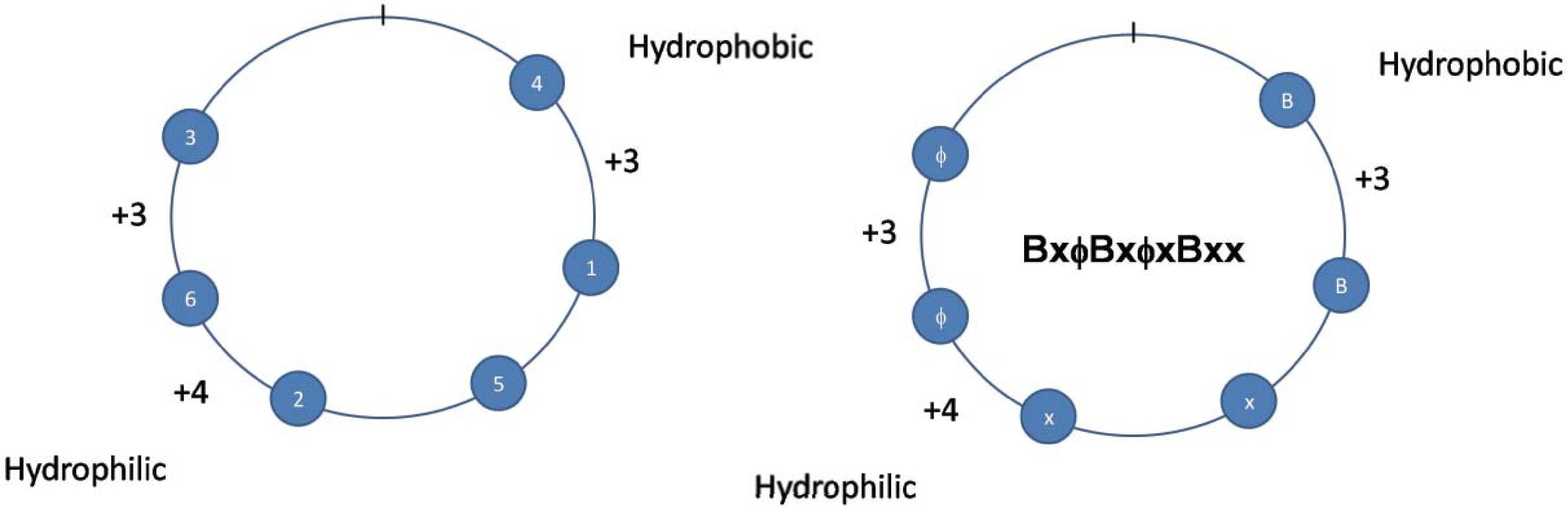


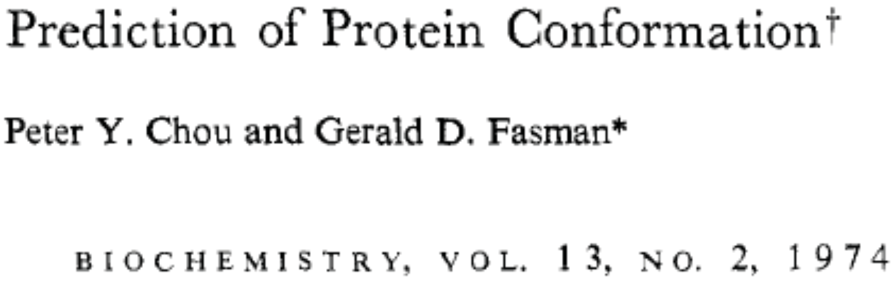
USE OF HELICAL WHEELS TO REPRESENT THE STRUCTURES OF PROTEINS AND TO IDENTIFY SEGMENTS WITH HELICAL POTENTIAL

MARIANNE SCHIFFER *and* ALLEN B. EDMUNDSON

From the Division of Biological and Medical Research, Argonne National Laboratory, Argonne, Illinois

Biophysical Journal Volume 7, Issue 2, March 1967, Pages 121–135


Hydrophobic
4
3
+3
+3
1
6
5
+4
2
Hydrophilic
Hydrophobic
B
ϕ
+3
+3
BxϕBxϕxBxx
B
ϕ
x
+4
x
Hydrophilic



Prediction of Protein Conformation†

Peter Y. Chou and Gerald D. Fasman*

BIOCHEMISTRY, VOL. 13, NO. 2, 1974

- Assembled statistical data from the small set of known structures
- Defined "propensity" for helix formation
- Crude rules to predict helical regions

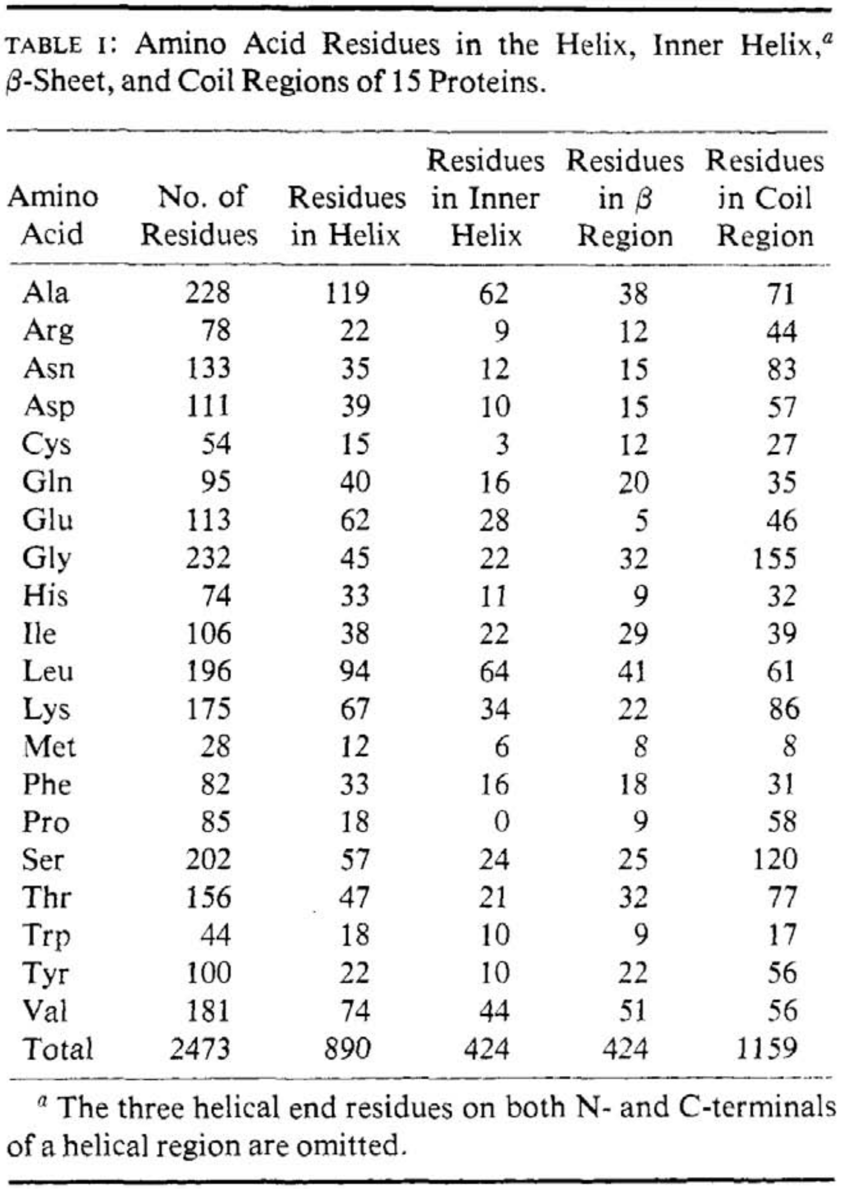
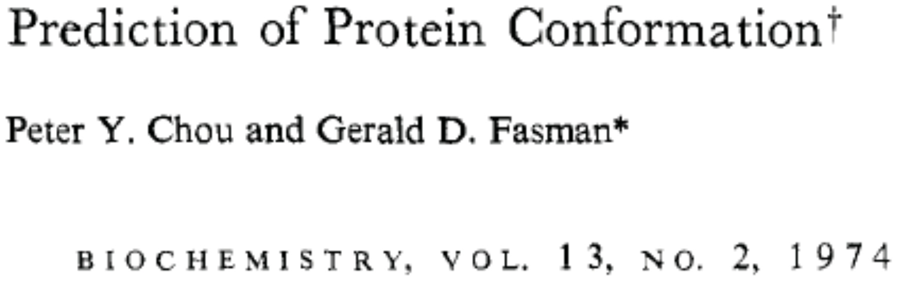


TABLE I: Amino Acid Residues in the Helix, Inner Helix,[a] β-Sheet, and Coil Regions of 15 Proteins.

Amino Acid	No. of Residues	Residues in Helix	Residues in Inner Helix	Residues in β Region	Residues in Coil Region
Ala	228	119	62	38	71
Arg	78	22	9	12	44
Asn	133	35	12	15	83
Asp	111	39	10	15	57
Cys	54	15	3	12	27
Gln	95	40	16	20	35
Glu	113	62	28	5	46
Gly	232	45	22	32	155
His	74	33	11	9	32
Ile	106	38	22	29	39
Leu	196	94	64	41	61
Lys	175	67	34	22	86
Met	28	12	6	8	8
Phe	82	33	16	18	31
Pro	85	18	0	9	58
Ser	202	57	24	25	120
Thr	156	47	21	32	77
Trp	44	18	10	9	17
Tyr	100	22	10	22	56
Val	181	74	44	51	56
Total	2473	890	424	424	1159

[a] The three helical end residues on both N- and C-terminals of a helical region are omitted.

Source: Chou, Peter Y., and Gerald D. Fasman. "Conformational Parameters for Amino Acidsin Helical, β-sheet, and Random Coil Regions Calculated from Proteins." *Biochemistry* 13, no. 2 (1974): 211-22.

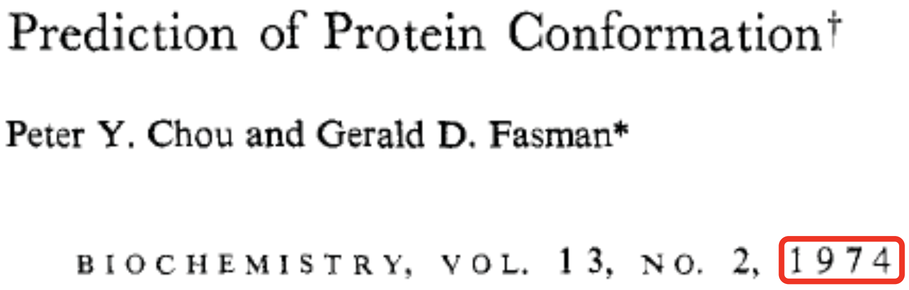


Prediction of Protein Conformation†

Peter Y. Chou and Gerald D. Fasman*

BIOCHEMISTRY, VOL. 13, NO. 2, 1974

- *Helix Nucleation*. Locate clusters of four out of six residues with a high propensity for forming helices.
- There are special cases for Asp and His which weakly nucleate and for Tyr, Asn, Pro and Gly which are considered helix breakers.
- *Helix Termination*. Extend the helical segment in *both* directions until terminated by tetrapeptides with low average helical propensity scores.
- Pro cannot occur in the alpha helix.

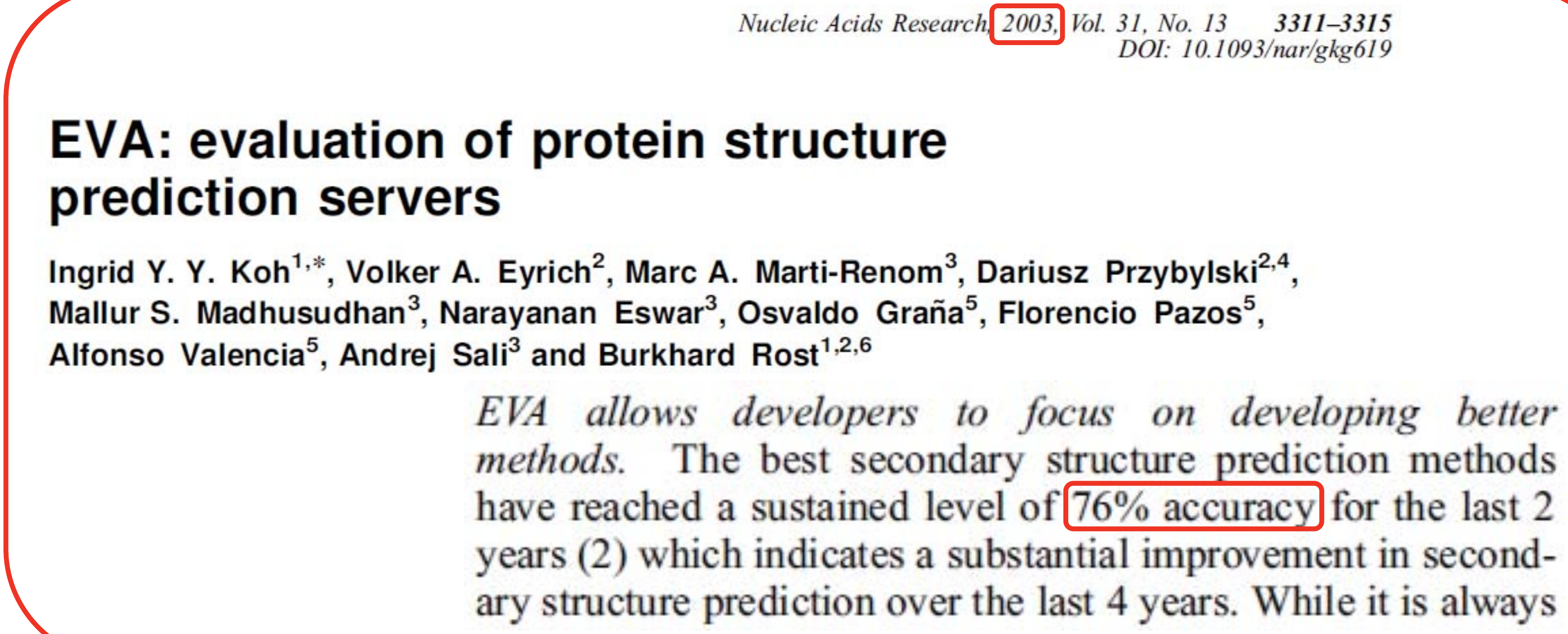


Prediction of Protein Conformation†

Peter Y. Chou and Gerald D. Fasman*

BIOCHEMISTRY, VOL. 13, NO. 2, 1974

~60% accuracy



Nucleic Acids Research, 2003, Vol. 31, No. 13 **3311–3315**
DOI: 10.1093/nar/gkg619

EVA: evaluation of protein structure prediction servers

Ingrid Y. Y. Koh[1,*], Volker A. Eyrich[2], Marc A. Marti-Renom[3], Dariusz Przybylski[2,4], Mallur S. Madhusudhan[3], Narayanan Eswar[3], Osvaldo Graña[5], Florencio Pazos[5], Alfonso Valencia[5], Andrej Sali[3] and Burkhard Rost[1,2,6]

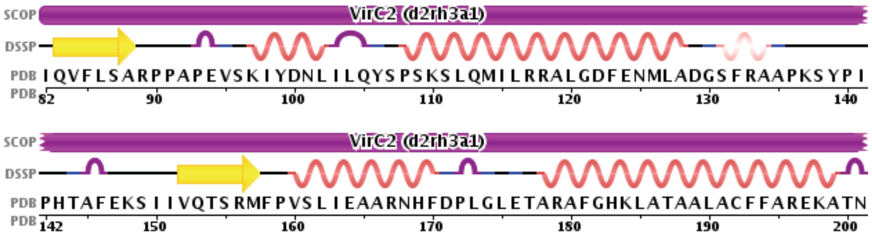
EVA allows developers to focus on developing better methods. The best secondary structure prediction methods have reached a sustained level of 76% accuracy for the last 2 years (2) which indicates a substantial improvement in secondary structure prediction over the last 4 years. While it is always

- Optional reading:
  - Chapter 12 of Zvelebil and Baum has an detailed description of current algorithms

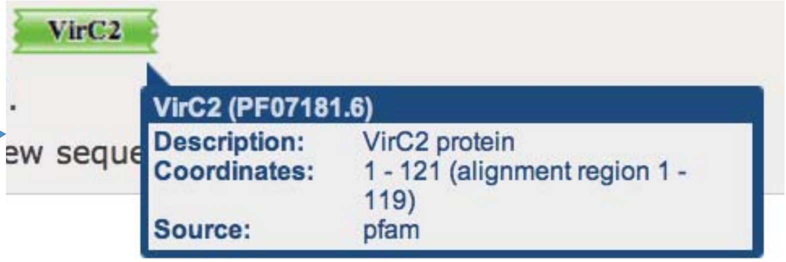
Other prediction problems

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

secondary structure


SCOP
VirC2 (d2rh3a1)
DSSP
PDB IQVFLSARPPAPEVSKIYDNLILQYSPSKSLQMILRRALGDFENMLADGSFRAAPKSYPI
PDB 82 90 100 110 120 130 140
SCOP
VirC2 (d2rh3a1)
DSSP
PDB PHTAFEKSIIVQTSRMFPVSLIEAARNHFDPLGLETARAFGHKLATAALACFFAREKATN
PDB 142 150 160 170 180 190 200

domain structure



VirC2 (PF07181.6)

Description: VirC2 protein

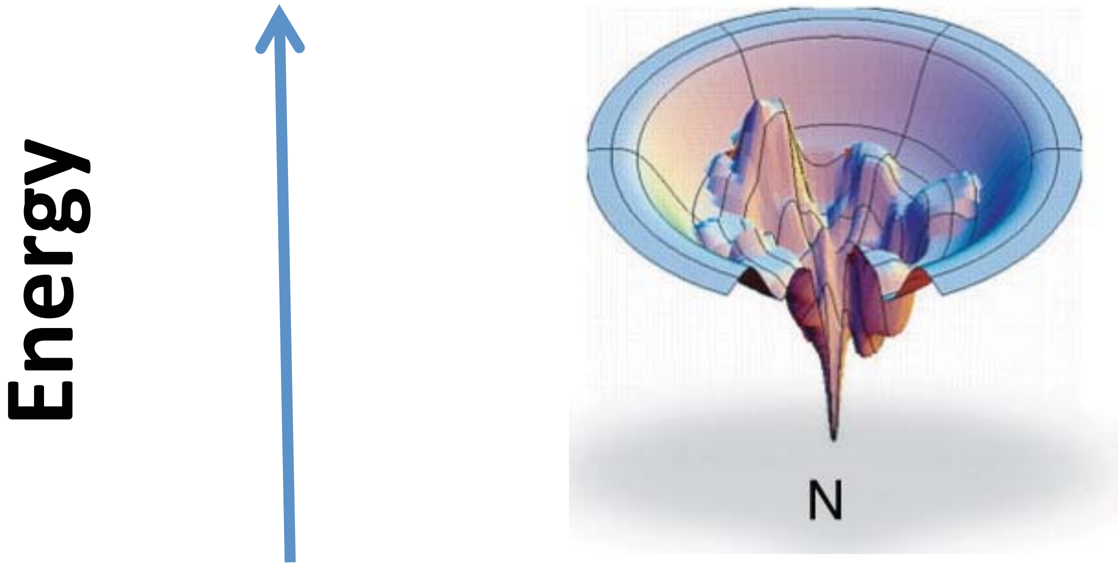
Coordinates: 1 - 121 (alignment region 1 - 119)

Source: pfam

novel 3D structure

Courtesy of RCSB Protein Data Bank.
Used with permission.

Computational Protein Folding



Courtesy of Nature Publishing Group. Used with permission.
Source: Dill, Ken A. and Hue Sun Chan. "From Levinthal to Pathways to Funnels."
Nature Structural Biology 4, no. 1 (1997): 10-9.

In principle, we don't even need a starting structure.

IQVFLSARPPAPEVSKIY
DNLILQYSPSKSLQMILR
RALGDFENMLADGSFR
AAPKSYPIPHTAFEKSIIV
QTSRMFPVSLIEAARN
HFDPLGLETARAFGHKL
ATAALACFFAREKATNS

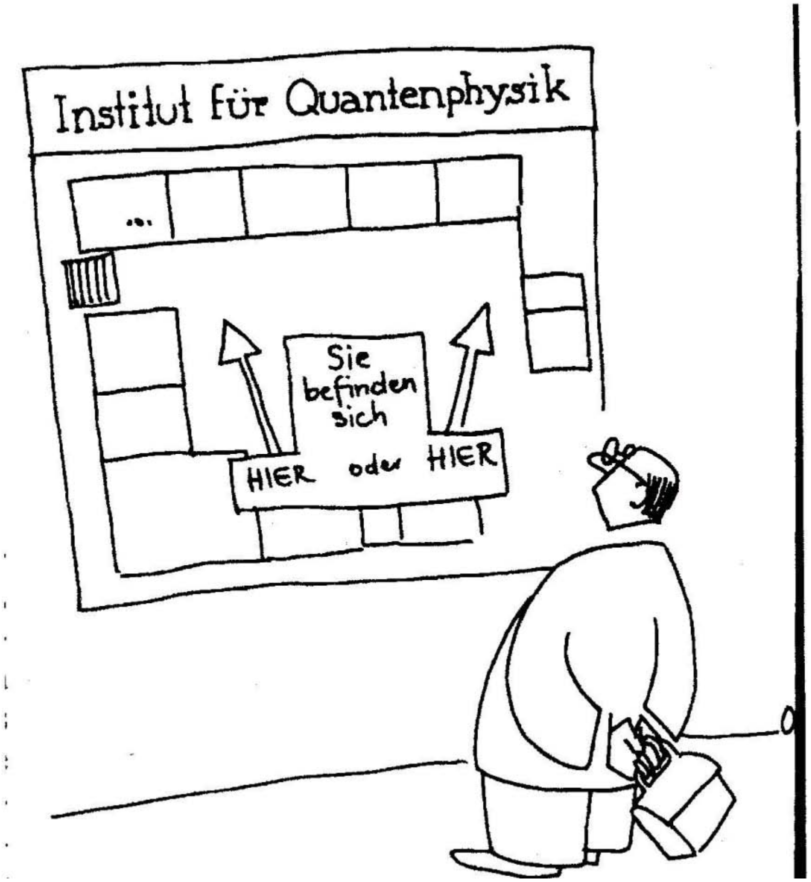
novel 3D structure

Courtesy of RCSB Protein Data Bank.
Used with permission.

Statisticians vs. Physicists

www.VADLO.com
"Data don't make any sense,
we will have to resort to statistics."

Courtesy of http://vadlo.com/. Used with permission.



Statisticians vs. Physicists

Rosetta

- Leverage everything we know about existing structures of proteins and peptides to build starting models
- Refine using a knowledge-based potential

DE Shaw

- DON'T CHEAT!
- Only use physical forces.
- Fold proteins by simulating the in vitro process