Bioinformatics Subtopics

Fold Recognition

Secondary Structure Prediction

Docking & Drug Design

Homology Modeling

Expression Clustering

Protein Geometry

Sequence Alignment

E-literature

Protein Flexibility

Gene Prediction

Database Design

Gene Annotation

Function Classification

Structure Classification

Large-Scale Genomic Surveys
Large-scale Information: GenBank Growth

<table>
<thead>
<tr>
<th>Year</th>
<th>Base Pairs</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>680338</td>
<td>606</td>
</tr>
<tr>
<td>1983</td>
<td>2274029</td>
<td>2427</td>
</tr>
<tr>
<td>1984</td>
<td>3368765</td>
<td>4175</td>
</tr>
<tr>
<td>1985</td>
<td>5204420</td>
<td>5700</td>
</tr>
<tr>
<td>1986</td>
<td>9615371</td>
<td>9978</td>
</tr>
<tr>
<td>1987</td>
<td>15514776</td>
<td>14584</td>
</tr>
<tr>
<td>1988</td>
<td>23800000</td>
<td>20579</td>
</tr>
<tr>
<td>1989</td>
<td>34762585</td>
<td>28791</td>
</tr>
<tr>
<td>1990</td>
<td>49179285</td>
<td>39533</td>
</tr>
<tr>
<td>1991</td>
<td>71947426</td>
<td>55627</td>
</tr>
<tr>
<td>1992</td>
<td>101008486</td>
<td>78608</td>
</tr>
<tr>
<td>1993</td>
<td>157152442</td>
<td>143492</td>
</tr>
<tr>
<td>1994</td>
<td>217102462</td>
<td>215273</td>
</tr>
<tr>
<td>1995</td>
<td>384939485</td>
<td>555694</td>
</tr>
<tr>
<td>1996</td>
<td>651972984</td>
<td>1021211</td>
</tr>
<tr>
<td>1997</td>
<td>1160300687</td>
<td>1765847</td>
</tr>
<tr>
<td>1998</td>
<td>2008761784</td>
<td>2837897</td>
</tr>
<tr>
<td>1999</td>
<td>3841163011</td>
<td>4864570</td>
</tr>
<tr>
<td>2000</td>
<td>8604221980</td>
<td>7077491</td>
</tr>
</tbody>
</table>
The next step:

What is a Conserved Domain?

Domains can be thought of as functional and/or structural units of a protein. There are two classifications: ordered versus disordered and whether or not the domain is strictly conserved across species. Typically domains are identified as recurring (sequence or structure) units, which may exist in various contexts. The image below illustrates 4 domains identified as structural units using the PDB database.

For this query sequence, the CD-Search service would identify the conserved domains indicated above. Click on the image below to launch the actual search. Good correspondence exists between structural units identified by purely geometric criteria and those assumed to be evolutionarily conserved. The region annotated as "Fumaric-like" was split into two by the Pfam domain parser.

Molecular evolution readily utilizes such domains as building blocks which may be recombined in different arrangements to mediate protein function. We define conserved domains as recurring units in molecular evolution whose extents can be determined by sequence and structure analysis.

Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The distinction between domains and motifs is not sharp, however, especially in the case of short repetitive units. Functional motifs are also present outside the scope of evolutionarily conserved domains. The CD database does not attempt to systematically collect these.
Ac-L-Ala-NHMe

Figure 7. Plot of main chain dihedral angles $\phi$ and $\psi$ (see Fig. 2 for definitions) experimentally determined for approximately 1000 non-glycine residues in eight proteins whose structures have been refined at high resolution (chosen to be representative of all categories of tertiary structure).

Figure 8. Plot of main chain dihedral angles $\phi$ and $\psi$ experimentally determined for the glycines in 20 high-resolution protein structures.

Ac-Gly-NHMe

Figure 2. Conformational energy contour map of $N$-acetyl-$N'$-methyl-glycinamide, for $\chi = 0^\circ$. Locations of minima are indicated by the filled circles (also see Table II). The contour lines are labeled with energy in kcal/mol above the minimum-energy point at (e.g. $\phi = 140^\circ$, $\psi = 30^\circ$).

Figure 3. Conformational energy contour map of $N$-acetyl-$N'$-methyl-glycinamide. Locations of minima are indicated by the filled circles (also see Table II). The contour lines are labeled with energy in kcal/mol above the minimum-energy point at (e.g. $\phi = 140^\circ$, $\psi = 30^\circ$) and ($\phi = 140^\circ$, $\psi = 70^\circ$).
Principal Protein Fold Classes

All alpha

All beta

alpha + beta

alpha / beta
Dynamic Programming Algorithm: Alternate Tracebacks

Correspond to Alternative Alignments

A B C N Y R Q C L C R P M
A Y C Y N R C K C R B P

A 8 7 6 6 5 4 4 3 3 2 1 0 0
Y 7 7 6 6 6 4 4 3 3 2 1 0 0
C 6 6 7 6 5 4 4 4 3 3 1 0 0
Y 6 6 6 5 6 4 4 3 3 2 1 0 0
N 5 5 5 6 5 4 4 3 3 2 1 0 0
R 4 4 4 4 5 4 3 3 2 2 0 0
C 3 3 4 3 3 3 3 4 3 3 1 0 0
K 3 3 3 3 3 3 3 3 3 2 1 0 0
C 2 2 3 2 2 2 2 3 2 3 1 0 0
R 2 1 1 1 1 2 1 1 1 1 2 0 0
B 1 2 1 1 1 1 1 1 1 1 1 0 0
P 0 0 0 0 0 0 0 0 0 0 0 1 0
Multiple Sequence Alignment (MSA)
COG 272, BRCT family

P. Bork et al
Community Assembly Through Adaptive Radiation in Hawaiian Spiders.

Phylogeny of spiny leg spider clade based on combined mitochondrial cytochrome oxidase I, 12S ribosomal DNA, and 16S ribosomal DNA sequences
1.6 Å electron density map
AV-COSY

Fig. 1. Stacked-plot representation of a symmetrized (Baumann et al., 1981) absolute value 500 MHz $^1$H COSY spectrum of a 0.02 m solution of BPT1 in a mixed solvent of 90% H$_2$O and 10% $^2$H$_2$O. pH 4.6, t = 80°C. The spectrum was recorded in ~24 h, the digital resolution is 3.3 Hz/point. The stacked plots afford a “3-dimensional” view of the spectrum. The 2 perpendicular frequency axes $\omega_1$ and $\omega_2$ are calibrated with the chemical shifts. Peaks corresponding to the one-dimensional spectrum are displayed on the diagonal from the upper right to the lower left corner, where some of the highest peaks have been truncated. “Cross peaks” manifesting J-connectivities between distinct lines on the diagonal are located in pairs, symmetrical with respect to the diagonal. Between 4-0 and 5-2 p.p.m., a band of artifactual peaks parallel to and in front of the diagonal spectrum are seen. These are a consequence of the water irradiation and the
Amide Proton Exchange Protection Factor
\[ F = \frac{k_{ex,\text{bound}}}{k_{ex,\text{free}}} \]
Enzyme Kinetics

A MODERN APPROACH

A. G. Morangoni

\[ v_o = \frac{k_{cat}}{K_M} [E][S] \]
Structural Validation of Homology

Adenylate Kinase

Guanylate Kinase

19% Seq ID

Z = 12.2
**Major Application:**

**Designing Drugs**

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).
Influenza A Virus (Bird Flu)
Non-Structural Protein 1 - Target for Antiviral Drug Design