MULTISEQ in VMD - Revealing How Nature Designs Proteins and RNAs

Luthey-Schulten Group

Department of Chemistry, Biophysics, and Beckman Institute
University of Illinois at Urbana-Champaign
Universal Phylogenetic Tree
three domains of life

Based on 16S rRNA

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

for review see Woese PNAS 2000

Woese, Olsen, Ibba, Soll MMBR 2000
After W. Doolittle, modified by G. Olsen
Phylogenetic Distributions

- Full Canonical
- Basal Canonical
- Non-canonical

increasing inter-domain of life Horizontal Gene Transfer

“HGT erodes the historical trace, but does not completely erase it….“ G. Olsen
Protein Structure Similarity Measure

**Q_H - Structural Homology**

fraction of native contacts for aligned residues + presence and perturbation of gaps

\[
Q_H = \kappa [q_{aln} + q_{gap}]
\]

\[
q_{aln} = \sum_{i<j-2} \exp \left[ -\frac{(r_{ij} - r_{i',j'})^2}{2\sigma_{ij}^2} \right]
\]

Structural Similarity Measure
the effect of insertions

“Gaps should count as a character but not dominate” C. Woese

\[ Q_H = 0.82 \]

\[ Q_H = 0.70 \]

\[ Q_H = 0.62 \]

\[
q_{gap} = \sum_{g_a} \sum_j \max \left\{ \exp \left[ -\frac{(r_{g_a,j} - r_{g_a,j'})^2}{2\sigma_{g_a,j}^2} \right] , \exp \left[ -\frac{(r_{g_a,j} - r_{g_a,j'})^2}{2\sigma_{g_a,j}^2} \right] \right\} \\
+ \sum_{g_b} \sum_j \max \left\{ \exp \left[ -\frac{(r_{g_b,j} - r_{g_b,j'})^2}{2\sigma_{g_b,j}^2} \right] , \exp \left[ -\frac{(r_{g_b,j} - r_{g_b,j'})^2}{2\sigma_{g_b,j}^2} \right] \right\}
\]
Protein structure encodes evolutionary information

sequence-based phylogeny

structure-based phylogeny

$\delta Q_H = 0.10$

JMB 2003, 2005

Da - AspRS archaeal genre

Db - AspRS bacterial genre
Protein structure reveals distant evolutionary events

Class I AARSs

Class II AARSs

Class I Lysyl-tRNA Synthetase

Class II Lysyl-tRNA Synthetase
Sequences define more recent evolutionary events

Conformational changes in the same protein.

ThrRS
T-AMP analog, 1.55 Å.  
T, 2.00 Å.

\( Q_H = 0.80 \)
Sequence identity = 1.00

ProRS
\textit{M. jannaschii}, 2.55 Å.  
\textit{M. thermoautotrophicus}, 3.20 Å.

\( Q_H = 0.89 \)
Sequence identity = 0.69

Structures for two different species.
Non-redundant Representative Sets

Too much information
129 Structures

Multidimensional QR factorization of alignment matrix, $A$.

$A = \begin{bmatrix} 
X & Y & Z & G \\
\end{bmatrix}$

Economy of information
16 representatives

QR computes a set of maximal linearly independent structures.


Numerical Encoding of Proteins in a Multiple Alignment

**Encoding Structure**
Rotated Cartesian + Gap = 4-space

- **Aligned position**
  \((x_{C_α}, y_{C_α}, z_{C_α}, 0)\)

- **Gapped position**
  \((0, 0, 0, g)\)

- **Gap Scaling**
  \(g = \frac{\|X\|_{F_4} + \|Y\|_{F_4} + \|Z\|_{F_4}}{\|G\|_{F_4}}\)

- adjustable parameter

**Sequence Space**
Orthogonal Encoding = 24-space

23 amino acids (20 + B, X, Z) + gap

- \(A = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\)
- \(B = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\)
- \(C = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)\)
- ...\)
- \(\text{GAP} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)\)

**Alignment is a Matrix with Linearly Dependent Columns**

\[
A = \begin{bmatrix}
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\end{bmatrix}
\]

\(m_{\text{aligned positions}} \rightarrow \text{n proteins} \rightarrow \text{encoded residue space} \rightarrow d=1, 2, 3, \ldots, N\)

\[
Q^T A_{(d)} P = Q^T (d) \begin{bmatrix}
X \\
Y \\
Z \\
G \\
\end{bmatrix}
\]

\(P = \tilde{R}_{(d)}\)

A maximal linearly independent subset can be determined with respect to a threshold, e.g., similarity measure threshold.
Class I AARSs

evolutionary events

5 Subclasses

Specificity – 11 Amino acids

Domain of life A,B,E
How many sequences are needed to represent the Subclass ILMV?

If each of ILMV was full canonical, then we would need $4 \times 3 = 12$ sequences.

Since M and V are basal, we need at least $2 \times 3 + 2 \times 2 = 10$ sequences.

We have 6 structures.
Evolutionary Profiles for Homology Recognition
AARS Subclass ILMV

The composition of the profile matters.
Choosing the right 10 sequence makes all the difference.

A. Sethi, P. O’Donoghue, Z. Luthey-Schulten (2005) JMB, PNAS
Genome Annotation

*M. jannaschii* genome was completely sequenced in 1996. Genome had four missing AARSs:

- AsnRS
- GlnRS

Indirect Mechanism

- LysRS Class I AARS
- CysRS ?

Cysteinyl-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J. Bacteriology*, Jan. 2004, **186**:8-14.

Ruan B, Nakano H, Tanaka M, Mills JA, DeVito JA, Min B, Low KB, Battista JR, and Söll D.

<table>
<thead>
<tr>
<th>Protein</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HisRS</td>
<td>1.1e-10</td>
</tr>
<tr>
<td>AspRS</td>
<td>1.9e-10</td>
</tr>
<tr>
<td>PheRS α-chain</td>
<td>9.5e-10</td>
</tr>
<tr>
<td>ThrRS</td>
<td>6.6e-04</td>
</tr>
<tr>
<td>ProRS</td>
<td>9.1e-03</td>
</tr>
<tr>
<td>SerRS</td>
<td>9.2e-03</td>
</tr>
<tr>
<td>putative CysRS</td>
<td>1.6e-02</td>
</tr>
<tr>
<td>AlaRS</td>
<td>5.1e-02</td>
</tr>
<tr>
<td>GlyRS</td>
<td>0.12</td>
</tr>
<tr>
<td>PheRS β-chain</td>
<td>0.15</td>
</tr>
<tr>
<td>DNA repair protein</td>
<td>7.5</td>
</tr>
</tbody>
</table>

*M. jannaschii* genome database search using EP of class II AARS with HMMER

MJ1660

Sethi, et. al., *PNAS*, **102**, 2005
Cysteine Biosynthesis in *Methanocaldococcus jannaschii*

Sauerwald et al. Science 2005
Evolutionary profile for HisA-HisF family


Sethi, et. al., PNAS, 2005.
Economy of Information
How many sequences are needed for profiles?

A. Sethi, P. O’Donoghue, ZLS, PNAS 102, 2005
Phylogenetic relationship between TIM barrels
Found in database search with HisA-HisF profile
Evolution of Structure and Function in AspRS

i) class II

ii) subclass IIB
   - anticodon binding (ACB) domain

iii) AspRS

iv) bacterial AspRS

v) E. coli AspRS

\[ \delta Q_H = 0.1 \]

<table>
<thead>
<tr>
<th>SCOP</th>
<th>QR order</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1b70a_</td>
<td>1</td>
</tr>
<tr>
<td>d1serb2</td>
<td>3</td>
</tr>
<tr>
<td>d1h4sb2</td>
<td>6</td>
</tr>
<tr>
<td>d1bbua2</td>
<td>4 ( #2 )</td>
</tr>
<tr>
<td>d1b8ab2</td>
<td>9 ( 5 4 )</td>
</tr>
<tr>
<td>d1n9wb2</td>
<td>10 ( 7 6 )</td>
</tr>
<tr>
<td>d1asza2</td>
<td>5 ( 3 3 )</td>
</tr>
<tr>
<td>d1sca2</td>
<td>7 ( 4 2 )</td>
</tr>
<tr>
<td>d1efwa3</td>
<td>8 ( 6 5 )</td>
</tr>
<tr>
<td>d1c0aa3</td>
<td>2 ( 1 1 )</td>
</tr>
</tbody>
</table>
Unifying the Worlds of Sequence and Structure
Multiseq in VMD: Merging the sequence and structure worlds

Version 1.83
2006 MultiSeq: New Features
Analyze the Evolution of Sequence and Structure
List of New Features in Multiseq

1. INPUT: Sequences and structures of proteins and nucleic acids from file or Blast searches of specialized databases:
   - Structural (PDB, SCOP, ASTRAL, NDB, VIPER..)
   - Sequence (NCBI, ASTRAL, modified tRNA, Viral)
   - Sequence Editor and Electronic Notebook

2. TOOLS:
   - Alignments (STAMP, CLUSTAL, TCoffee)
   - Database Searches - BLAST and VMD/Multiple DB searches
   - QR reduction, Phylogenetic tree - UPGMA, NJ
   - Conservation Mappings, RMSD plots
   - Covariance and Coordination Analysis
Acknowledgements

Patrick O’Donoghue
Anurag Sethi

Rommie Amaro
Felix Autenrieth
Alexis Black

John Eargle
Corey Hardin
Taras Pogorelov

Elijah Roberts
Dan Wright

Graphics Programmers VMD
Elijah Roberts, Dan Wright, John Eargle
John Stone

Collaborators
Evolutionary Studies
Gary Olsen, Carl Woese (UIUC)

QR Algorithms
Mike Heath (UIUC)

Protein Structure Prediction
Peter Wolynes, Jose Onuchic (UCSD)
Ken Suslick (UIUC)

Funding
NSF, NIH
Demonstration of New Multiseq Features

1. AspRS structures: STAMP multiple structure alignment. Color by structure (Qpair) and sequence conservation. Tcl script - seq ID and Sec. Str. Information in beta field.

2. Sequence Editor and Electronic Notebook


4. Phylogenetic trees of structure and sequences: HGT and QR algorithm for sequences. Evolutionary profiles