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

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

$$Q_H = \sum \exp \left[ -\frac{(r_{ij} - r_{i',j'})^2}{2\sigma_{ij}^2} \right]$$

Structural Similarity Measure
the effect of insertions

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

\[ Q_H = \begin{cases} 0.82 \\
0.70 \\
0.62 \end{cases} \]

\[ q_{gap} = \sum_{g_a} \sum_{j} \max \left\{ \exp \left[ \frac{-\left(r_{g_a,j} - r_{g_a,j'}\right)^2}{2\sigma_{g_a,j}^2} \right], \exp \left[ \frac{-\left(r_{g_a,j} - r_{g_a,j'}\right)^2}{2\sigma_{g_a,j}^2} \right] \right\} \]

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

sequence-based phylogeny

structure-based phylogeny

JMB 2005
MMBR 2003

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

structure-based phylogenetics

sequence-structure overlap
Sequences define more recent evolutionary events

Conformational changes in the same protein.

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

$Q_H = 0.80$
Sequence identity = 1.00

ProRS
*M. jannaschii*, 2.55 A.
*M. thermoautotrophicus*, 3.20 A.

$Q_H = 0.89$
Sequence identity = 0.69

Structures for two different species.
Non-redundant Representative Sets

Too much information
129 Structures

Multidimensional QR factorization of alignment matrix, $A$.

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

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**
\((x_{C\alpha}, y_{C\alpha}, z_{C\alpha}, 0)\)

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

**Gap Scaling**
\[ g = \frac{\|X\|_F^4 + \|Y\|_F^4 + \|Z\|_F^4}{\|G\|_F^4} \]

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

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

\[ A = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \]
\[ B = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \]
\[ C = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \]
\[ \ldots \]
\[ GAP = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1) \]

Alignment is a Matrix with Linearly Dependent Columns

\[ A = \]
\[ \begin{array}{cccc}
\text{d=1} & \text{d=2} & \text{d=3} & \text{d=N} \\
\text{m aligned positions} & \text{n proteins} & \text{encoded residue space} & \\
\end{array} \]

\[ Q^T_{(d)} A_{(d)} P = Q^T_{(d)} \]

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
Profile of the ILMV Subclass

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.
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
- GluRS

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

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

\[
\begin{align*}
\text{O-phosphoserine} & \xrightarrow{\text{MJ1594 (SerB)}} \text{L-serine} \\
\text{COOH} & \quad \text{H–C–CH}^2\text{–}\text{OH} + \text{PO}_4^{3-} \\
\text{NH}_2 & \quad \text{NH}_2
\end{align*}
\]

\[
\begin{align*}
\text{COO} & \quad \text{H–C–CH}^2\text{–}\text{OP}O_3H_2 \\
\text{3' tRNA}^{\text{Cys}} & \quad \text{3' tRNA}^{\text{Cys}} \\
\text{NH}_2 & \quad \text{NH}_2 \\
\text{Sep-tRNA}^{\text{Cys}} & \quad \text{Sep-tRNA}^{\text{Cys}} \\
\text{AMP + PP}_i & \quad \text{S source} \\
\text{MJ1678 (SepCysS)} & \quad \text{Cys-tRNA}^{\text{Cys}} \\
\text{COO} & \quad \text{H–C–CH}^2\text{–}\text{SH} + \text{PO}_4^{3-} \\
\text{3' tRNA}^{\text{Cys}} & \quad \text{Cys-tRNA}^{\text{Cys}}
\end{align*}
\]

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?

<table>
<thead>
<tr>
<th>Profile</th>
<th>N_{Seq}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam full</td>
<td>250</td>
</tr>
<tr>
<td>Pfam seed</td>
<td>147</td>
</tr>
<tr>
<td>Pfam/QR 40%</td>
<td>21</td>
</tr>
<tr>
<td>QR 15%</td>
<td>2</td>
</tr>
<tr>
<td>QR 30%</td>
<td>4</td>
</tr>
<tr>
<td>QR 40%</td>
<td>13</td>
</tr>
<tr>
<td>QR 100%</td>
<td>238</td>
</tr>
</tbody>
</table>

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

δQ_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</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>d11sca2</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

Eliminate Redundancy

Plus More Functions
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

Funding
NSF, NIH

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)
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
1. Show distance matrix for NJ/UPGMA for small number 3-4 sequences. Give algebraic equations needed for NJ.

2. MP/ML trees: Animate through several tree topologies generated by paup to describe the search through tree space.
Maximum Parsimony
Fitch optimization

Assign characters to the ancestral nodes and calculate the number of steps (sequence changes) required by a data set on a given tree.

“Downpass” algorithm traces back through the tree from leaves to root.

If descendent characters intersect add 0 to total length.

If descendent characters do not intersect, their union set is assigned to the node add 1 to total length.

The intersection of C and (C,T) is C. This ancestral node is assigned the “state” C. The total length is unchanged.

The length on this tree for this site is 2. The length of this topology for the sequences in the alignment is the sum of length over all sites gives.

Adapted from Mark E. Siddall, http://research.amnh.org/~siddall/methods/day3.html