Introduction to evolutionary concepts and VMD/MultiSeq - Part I

Characterizing your systems

Zaida (Zan) Luthey-Schulten
Dept. Chemistry, Physics, Beckman Institute, Institute of Genomics Biology, & Center for Biophysics

Workshop April 2015, UIUC
NIH Center Macromolecular Modeling and Bioinformatics
VMD/MultiSeq - “A Tool to Think”

Carl Woese - “VMD is far from a simple visualization tool for a biologist, it is a true thinking tool. Without it a whole class of biological hypotheses would simply not exist.”

UPT - Woese 16S rRNA

Evolutionary profiles for protein structure & function prediction

Signatures ribosomal evolution

LSU (23S rRNA + rproteins)
Why Look at More Than One Sequence?

1. Multiple Sequence Alignment shows patterns of conservation

2. Are these positions functionally important? Active sites, folding,..

3. What and how many sequences should be included?

4. Where do I find the sequences and structures for MS alignment?

5. How to generate pairwise and multiple sequence alignments?
New Tools in VMD/MultiSeq

Protein / RNA Sequence Data

SwissProt DB (400K), Greengenes RNA (100K) Signatures, Zoom

Metadata Information, Clustal, MAFFT & Phylogenetic Trees

RAXml Trees, Genomic Content, Temperature DB

Blast & PsiBlast

Sequence Editor

Sequence /Structure Alignment

Protein & RNA secondary structure

QR non-redundant seq / str sets

Cluster analysis / Bioinformatics scripting

Tutorials MultiSeq/AARS

EF-Tu/Ribosome

Protein (RNA) Folding, Structure, & Function
Protein:RNA Complexes in Translation
Evolutionary Analysis & Dynamics

“Evolution AARS Structure” MMBR 2003
“Evol. Profiles Class I&II AARS” JMB 2005
“Evolution SepRS/CysRS” PNAS 2005
“Dynamic Signaling Network” PNAS 2009
“Exit Strategy Charged tRNA” JMB 2010
“Mistranslation in Mycoplasma” PNAS 2011
“Capture and Selection of ATP” JACS 2013

“Dynamical Recognition Novel Amino Acids” JMB 2008
“tRNA Dynamics” FEBS 2010

r-Proteins/r-RNA
Ribosome LSU
“Motion L1 Stalk:tRNA” JMB 2010,
“Ribosome Biogenesis” JPC 2012,3
“Whole cell simulations on GPUs“ IEEE 2009, Plos CB 2011, PRL 2011,
Basic principles of evolutionary analysis for proteins & RNAs

• Comparative analysis of sequences and structures
• Multiple sequence alignments (gaps and editing)
• Sequence and structure phylogenetic trees*
• Reference to 16S rRNA tree
• Horizontal or lateral gene transfer events
• Genomic context
• Evolutionary profiles representing diversity
• Conservation analysis of evolutionary profiles

*Various models of evolutionary change
Alignment of ~200 EF-Tu sequences in VMD/MultiSeq

ClustalW alignment
~ 5 minutes

MAFFT* alignment
~ 30 seconds
More sequences!

* “Mafft” Katoh, Misawa, Kuma, Miyata, NAR 2002, 2005

“G” scattered around gaps

“G” aligned
Sequence Alignment & Dynamic Programming

Seq. 1: $a_1 \ a_2 \ a_3 \ - \ - \ a_4 \ a_5 \ldots \ a_n$

Seq. 2: $c_1 \ - \ c_2 \ c_3 \ c_4 \ c_5 \ - \ldots \ c_m$

Needleman-Wunsch alignment algorithm

$$H(i, j) = \text{MAX} \begin{cases} H(i-1, j-1) + S[a(i), b(j)] \\ H(i, j-k) - W(k), \\ H(i-m, j) - W(m) \end{cases}$$

$S$: substitution matrix

Score Matrix H: Traceback

gap penalty $W = -6$

Needleman-Wunsch Global Alignment

### Similarity Values

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>G</th>
<th>K</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>5</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>6</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>P</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>7</td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>-2</td>
<td>5</td>
<td>-1</td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>-2</td>
<td>5</td>
<td>-1</td>
</tr>
<tr>
<td>P</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>7</td>
</tr>
</tbody>
</table>

### Initialization of Gap Penalties

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>G</th>
<th>K</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>-6</td>
<td>5</td>
<td>-3</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-12</td>
<td>-3</td>
<td>6</td>
<td>-2</td>
</tr>
<tr>
<td>P</td>
<td>-18</td>
<td>-2</td>
<td>-2</td>
<td>7</td>
</tr>
<tr>
<td>K</td>
<td>-24</td>
<td>-1</td>
<td>-2</td>
<td>5</td>
</tr>
<tr>
<td>K</td>
<td>-30</td>
<td>-1</td>
<td>-2</td>
<td>5</td>
</tr>
<tr>
<td>P</td>
<td>-36</td>
<td>-2</td>
<td>-2</td>
<td>7</td>
</tr>
</tbody>
</table>
Filling out the Score Matrix H

![Score Matrix Diagram]

http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html
## Traceback and Alignment

![Alignment Diagram](image)

### The Alignment

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>G</th>
<th></th>
<th>K</th>
<th></th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>-6</td>
<td>-6</td>
<td>-12</td>
<td>-1</td>
<td>-13</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-12</td>
<td>-1</td>
<td>-13</td>
<td>11</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-18</td>
<td>-7</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-24</td>
<td>-13</td>
<td>-1</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-30</td>
<td>-18</td>
<td>-7</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-36</td>
<td>-25</td>
<td>-13</td>
<td>-2</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

Traceback (blue) from optimal score

http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html
STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs
   - Multiple Sequence alignment
   - Ridged Body “Scan”
   - Pairwise Alignments and Hierarchical Clustering

2. Refine Initial Alignment & Produce Multiple Structural Alignment

\[ P_{ij} = \left( e^{-\frac{d_{ij}^2}{2E_1}} \right) \left( e^{-\frac{s_{ij}^2}{2E_2}} \right) \]

probability that residue i in structure A is equivalent to residue j in structure B.

\[ d_{ij} \text{ — distance between } i \text{ & } j \]

\[ s_{ij} \text{ — conformational similarity; function of RMS between } i-1, i, i+1 \text{ and } j-1, j, j+1. \]

- Dynamic Programming (Smith-Waterman) through P matrix gives optimal set of equivalent residues.
- This set is used to re-superpose the two chains. Then iterate until alignment score is unchanged.
- This procedure is performed for all pairs with no gap penalty

Multiple Structural Alignments

STAMP – cont’d

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

\[
S_c = \frac{S_p}{L_p} \left( \frac{L_p - i_A}{L_A} - \frac{L_p - i_B}{L_B} \right)
\]

\[
S_p = \sum_{\text{align. path}} P_{ij}
\]

- \(L_p, L_A, L_B\) — length of alignment, sequence A, sequence B
- \(i_A, i_B\) — length of gaps in A and B.

Multiple Alignment:
- Create a dendrogram using the alignment score.
- Successively align groups of proteins (from branch tips to root).
- When 2 or more sequences are in a group, then average coordinates are used.
Structural Overlaps - STAMP

Ribosome large subunit showing ribosomal proteins L2 and L3

180,000 atoms in 4 rRNAs and 58 proteins
Universal Phylogenetic Tree

3 domains of life

Reference 16S rRNA tree

Leucyl-tRNA synthetase displays the full canonical phylogenetic distribution.

For review see Woese *PNAS* 2000

Woese, Olsen, Ibba, Soll *MMBR* 2000
Look for horizontal gene transfer events
increasing inter-domain of life  
Horizontal Gene Transfer

“HGT erodes the historical trace, but does not completely erase it....” G. Olsen

Woese, Olsen, Ibba, Soll MMBR 2000
Protein Structure Similarity Measure

**Q\textsubscript{H} Structural Homology**
fraction of native contacts for aligned residues +
presence and perturbation of gaps

\[ Q_H = \sum \left[ q_{aln} + q_{gap} \right] \]

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

Structural Similarity Measure: The effect of insertions

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

\[ Q_H = 0.82 \quad 0.70 \quad 0.62 \]

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

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

sequence-based phylogeny

structure-based phylogeny

Euryarchaeota

Crenarchaeota Thermoprotei

Deinococcus-Thermus 2*

Metazoa/Fungi

Euryarchaeota Halobacteria

AsnRS

Deinococcus-Thermus 1

Firmicutes Mollicutes

Firmicutes Bacilli

Firmicutes Clostridia

Bacteroidetes

γ-Proteobacteria

β-Proteobacteria

Cyanobacteria

ε-Proteobacteria

Chlamydiae

Thermotogae

Aquificae

Spirochaetes

Actinobacteria

Chlorobi

α-Proteobacteria

P. kodakaraensis d1b8aa2

T. thermophilus d1n9wb2*

Metazoa/Fungi

S. cerevisiae d1asza2

AsnRS T. thermophilus d11sca2

Deinococcus-Thermus 1

T. thermophilus d1efwa3

γ-Proteobacteria

E. coli d1c0aa3

20 changes

δQ_H = 0.10

bacterial insertions

archaeal helix extensions, insertion

Da - AspRS archaeal genre

Db - AspRS bacterial genre
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 event:

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.
Relationship Between Sequence & Structure

The sequence signal degrades rapidly.

sequence identity < 10%


Structural superposition of AlaRS & AspRS.

Sequence id = 0.055, $Q_H = 0.48$

Structural alignment & visualization software MultiSeq/VMD
Non-redundant Representative Profiles

Too much information
129 Structures

Economy of information
16 representatives

Multidimensional QR factorization of alignment matrix, \( A \).

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

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**
  \(g = \gamma \left( \frac{\mathcal{F}_4}{\mathcal{F}_4} \right) = \gamma \left( \frac{\|X\|_4 + \|Y\|_4 + \|Z\|_4}{\|G\|_4} \right)\)

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

- 23 amino acids \((20 + B, X, Z) + \text{gap}\)

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

Alignment is a Matrix with Linearly Dependent Columns

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

\(Q_{(d)}^T A_{(d)} P = Q_{(d)}^T\)

A maximal linearly independent subset can be determined with respect to a threshold, e.g., similarity measure threshold.
Design - Evolution of Structure and Function in Class II

i) class II

ii) subclass IIB

- anticodon binding (ACB) domain

iii) AspRS

- E. coli

- bacterial insert domain

iv) bacterial AspRS

v) E. coli AspRS

JMB 2005

δQ_H = 0.1

P. kodakaraensis

T. thermophilus 2*

S. cerevisiae

T. thermophilus 1

E. coli

Fb T. thermophilus

S T. thermophilus

Pa T. thermophilus

K2 E. coli

Da

N T. thermophilus

T. thermophilus 1

bacterial insert

E. coli

SCOP

d1b70a_ 1

d1serb2 3

d1h4sb2 6

d1bbua2 4

d1b8ab2 9

d1n9wb2 10

d1asza2 5

d11sca2 7

d1efwa3 8

d1c0aa3 2

QR order

①

③

⑥

②

⑨

⑩

③

②

⑧

②
Summary Structural Evolutionary Profiles

1. Structures often more conserved than sequences!! Similar structures at the Family and Superfamily levels. Add more structural information to identify core and variable regions

2. Which structures and sequences to include? Use evolution and eliminate redundancy with QR factorization
What is MultiSeq?

• MultiSeq is an extension to VMD that provides an environment to combine sequence and structure data
• A platform for performing bioinformatics analyses within the framework of evolution
• Provides software for improving the signal-to-noise ratio in an evolutionary analysis by eliminating redundancy (StructQR, SeqQR, Evolutionary Profiles “EP”)
• Visualizes computationally derived metrics ($Q_{res}$, $Q_H$,..) or imported experimental properties

• Integrates popular bioinformatics tools along with new algorithms (ClustalW, MAFFT, BLAST, STAMP, Signatures, Mutual information, QR, PT,....)
New Tools in VMD/MultiSeq

Protein / RNA Sequence Data

SwissProt DB (400K), Greengenes RNA (100K) Signatures, Zoom

Metadata Information, Clustal & Phylogenetic Trees

RAXml Trees, Genomic Content, Temperature DB

Blast & PsiBlast

Sequence Editor

Sequence /Structure Alignment

Protein & RNA secondary structure

QR non-redundant seq / str sets

Cluster analysis / Bioinformatics scripting

Tutorials MultiSeq/AARS

EF-Tu/Ribosome

MultiSeq Combines Sequence and Structure

- Align sequences or structures; manually edit alignments
- View data colored by numerous metrics including structural conservation and sequence similarity
- Synchronized coloring between 1D and 3D views
# Load large sequence sets

<table>
<thead>
<tr>
<th>Swiss-Prot (Proteins)</th>
<th>Greengenes (RNA)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curated sequences</td>
<td>Environmental 16S rRNA</td>
</tr>
<tr>
<td>392,667 sequences</td>
<td>90,654 entries</td>
</tr>
<tr>
<td>Unaligned</td>
<td>Aligned (7682 positions)</td>
</tr>
<tr>
<td>177 MB on disk</td>
<td>670 MB on disk</td>
</tr>
<tr>
<td>2 minutes to load</td>
<td>2.5 minutes to load *</td>
</tr>
<tr>
<td>2.4 GB memory used</td>
<td>4.0 GB memory used*</td>
</tr>
</tbody>
</table>
Sequence editor

• New sequence API allows editing of large alignments. Align closely related sequences by group, combine groups, and then manually correct.
• Zoom window gives an overview of the alignment, quickly move the editing window to any part of the alignment.

660 sequences of ribosomal protein S4 from all complete bacterial genomes.*

Phylogenetic tree editor

- Automatically add annotations and colors to phylogenetic trees based on taxonomy, enzyme, temperature class, and/or MultiSeq groupings.

A cluster of five proteobacterial sequences branch near the cyanobacterial sequences. These are cases of horizontal gene transfer.

Maximum likelihood tree of 660 S4 sequences reconstructed using RAxML.

Elijah Roberts 2009
Edit the physical layout of the tree

- Nodes with low support can be removed.
- Nodes can be rotated for easier reading.
Phylogenetic tree generation

- Generate distance based trees only over well-aligned columns (no indels).
- Export alignments in Phylip format (PHY) compatible with RAxML for maximum likelihood reconstructions.
- Import Newick trees from phylogenetic reconstruction programs (including RAxML).
Scripting MultiSeq

• All MultiSeq functions can now be scripted.
• Scripting an analysis provides benefits:
  – It can be checked for correctness.
  – It can be quickly repeated by anyone.
  – It can be modified later with new functionality.
  – It can be run on a cluster in VMD text mode.
    (if it can be easily broken into independent chunks)
• Many functions are too user specific and/or too complex to be turned into a GUI.
• Some examples of MultiSeq scripts…
Genome content

- When using sequence from fully sequenced genomes, additional information is available in the genome content.
- Conservation of gene ordering, neighbors, or intergenic regions can provide additional evolutionary information not contained in the sequence.
- Gene names and ordering can be obtained from the genome PTT files, want to organize the information in an evolutionarily meaningful manner.

<table>
<thead>
<tr>
<th>Location</th>
<th>Strand</th>
<th>Length</th>
<th>PID</th>
<th>Gene</th>
<th>Synonym</th>
<th>Code</th>
<th>COG</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>3437638..3438021</td>
<td>-</td>
<td>127</td>
<td>16131173</td>
<td>rplQ</td>
<td>b3294</td>
<td>-</td>
<td>COG0203J</td>
<td>50S ribosomal subunit protein L17</td>
</tr>
<tr>
<td>3438062..3439051</td>
<td>-</td>
<td>329</td>
<td>16131174</td>
<td>rpoA</td>
<td>b3295</td>
<td>-</td>
<td>COG0202K</td>
<td>RNA polymerase, alpha subunit</td>
</tr>
<tr>
<td>3439077..3439697</td>
<td>-</td>
<td>206</td>
<td>16131175</td>
<td>rpsD</td>
<td>b3296</td>
<td>-</td>
<td>COG0522J</td>
<td>30S ribosomal subunit protein S4</td>
</tr>
<tr>
<td>3439731..3440120</td>
<td>-</td>
<td>129</td>
<td>16131176</td>
<td>rpsK</td>
<td>b3297</td>
<td>-</td>
<td>COG0100J</td>
<td>30S ribosomal subunit protein S11</td>
</tr>
<tr>
<td>3440137..3440493</td>
<td>-</td>
<td>118</td>
<td>16131177</td>
<td>rpsM</td>
<td>b3298</td>
<td>-</td>
<td>COG0099J</td>
<td>30S ribosomal subunit protein S13</td>
</tr>
<tr>
<td>3440640..3440756</td>
<td>-</td>
<td>38</td>
<td>16131178</td>
<td>rpmJ</td>
<td>b3299</td>
<td>-</td>
<td>COG0257J</td>
<td>50S ribosomal subunit protein L36</td>
</tr>
<tr>
<td>3440788..3442119</td>
<td>-</td>
<td>443</td>
<td>16131179</td>
<td>secY</td>
<td>b3300</td>
<td>-</td>
<td>COG0201U</td>
<td>preprotein translocase membrane subunit</td>
</tr>
<tr>
<td>3442127..3442561</td>
<td>-</td>
<td>144</td>
<td>16131180</td>
<td>rplO</td>
<td>b3301</td>
<td>-</td>
<td>COG0200J</td>
<td>50S ribosomal subunit protein L15</td>
</tr>
<tr>
<td>3442565..3442744</td>
<td>-</td>
<td>59</td>
<td>16131181</td>
<td>rpmD</td>
<td>b3302</td>
<td>-</td>
<td>COG1841J</td>
<td>50S ribosomal subunit protein L30</td>
</tr>
<tr>
<td>3442748..3443251</td>
<td>-</td>
<td>167</td>
<td>16131182</td>
<td>rpsE</td>
<td>b3303</td>
<td>-</td>
<td>COG0098J</td>
<td>30S ribosomal subunit protein S5</td>
</tr>
</tbody>
</table>
Combined genomic context/phylogenetic tree

• Use a script to walk through a phylogenetic tree, find the genome content near the source gene, create a graphical representation of the combined data.

```plaintext
proc draw_genome_context_of_phylogeny {args} {

# Load the sequences.
set alignment [::SeqData::Fasta::loadSequences $alignmentFilename]

# Load the tree
set tree [::PhyloTree::Newick::loadTreeFile $treeFilename]

# Reorder the alignment by the tree.
set treeAlignment {}
set leafNodes [::PhyloTree::Data::getLeafNodes $tree]
foreach node $leafNodes {
    set foundNode 0
    set nodeName [::PhyloTree::Data::getNodeName $tree $node]
    foreach sequence $alignment {
        if {$nodeName == [::SeqData::getName $sequence]} {
            lappend treeAlignment $sequence
            set foundNode 1
            break
        }
    }
}

# Draw the genomic context.
drawGenomicContextOfAlignment $outputFilename $treeAlignment $contextDistance $scaling $genomeDirectory
}
```
Combined genomic context/phylogenetic tree

proc drawGenomicContextOfAlignment {outputFilename alignment contextDistance scaling genomeDirectory} {
    foreach sequence $alignment {
        # Make sure we have the GI number for this sequence.
        set giNumber [::SeqData::getSourceData $sequence "gi"]
        # Make sure we can tell which genome this sequence is from.
        set taxonomy [join [::SeqData::getLineage $sequence 1 0 1] "," ]
        if {![info exists genomeTaxonomyMap($taxonomy)]} {
            error "ERROR) Unknown genome for sequence [::SeqData::getName $sequence]: $taxonomy"
        }
        # Go through each of the genome context files for the genome.
        set foundGene 0
        foreach genomeName $genomeTaxonomyMap($taxonomy) {
            ...
        }
        # Draw the genomic context.
        drawMultipleGenomicContext $outputFilename $alignment $geneFiles $genePositions $geneStrands $contextDistance
    }
}

Betaproteobacteria, Thiobacillus denitrificans ATCC 25259
Betaproteobacteria, Azoarcus sp. BH72
Betaproteobacteria, Azoarcus sp. EbN1
Betaproteobacteria, Dechloromonas aromatica RC8
Betaproteobacteria, Nitrosospira multiformis ATCC 25196
Betaproteobacteria, Nitrosomonas europaeanus C91
Gammaproteobacteria, Psychrobacter arcticus 273-
Gammaproteobacteria, Psychrobacter cryohalolentis K
Gammaproteobacteria, Psychrobacter sp. PRef-1
Gammaproteobacteria, Acinetobacter sp. ADP1
Gammaproteobacteria, Acinetobacter baumannii SDF
Gammaproteobacteria, Acinetobacter baumannii AYE
Gammaproteobacteria, Acinetobacter baumannii ACICU
Genome content future directions

- Genome content still a work in progress.
- Good candidate for a GUI: combined phylogenetic tree/genome content viewer.
- Can also use COG codes to color by gene function.
- Still need API for manipulating PTT files.


Genome content of ribosomal protein S4 by occurrence of the gene in the alpha operon.

Fifteen Clostridia genomes contain two copies of S4: one zinc-binding and one zinc-free.
Flexible Grouping of Data

- Automatically group data by taxonomic classification to assist in evolutionary analysis (HGT) or create custom groups
- Apply metrics to groups independently, e.g. bacterial signal

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Taxonomic Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1asy_A</td>
<td>Eukaryota:Fungi</td>
</tr>
<tr>
<td>1eov_A</td>
<td>Eukaryota:Fungi</td>
</tr>
<tr>
<td>SYDC_YEAST</td>
<td>Eukaryota:Fungi</td>
</tr>
<tr>
<td></td>
<td>Eukaryota:Metazoa</td>
</tr>
<tr>
<td>SYD_CAEL</td>
<td>Eukaryota:Metazoa</td>
</tr>
<tr>
<td>SYD_HUMAN</td>
<td>Eukaryota:Metazoa</td>
</tr>
<tr>
<td>SYD_MOUSE</td>
<td>Eukaryota:Metazoa</td>
</tr>
<tr>
<td></td>
<td>Archaea:Crenarchaeota</td>
</tr>
<tr>
<td>SYD_AERFE</td>
<td>Archaea:Crenarchaeota</td>
</tr>
<tr>
<td></td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>1n9w_A</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>1b8a_A</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>SYD_METMA</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>SYD_HALN1</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>SYD_THEAC</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td>SYD_PYRHO</td>
<td>Archaea:Euryarchaeota</td>
</tr>
<tr>
<td></td>
<td>Bacteria:Proteobacteria</td>
</tr>
<tr>
<td>1I0w_A</td>
<td>Bacteria:Proteobacteria</td>
</tr>
<tr>
<td>1I2_A</td>
<td>Bacteria:Proteobacteria</td>
</tr>
</tbody>
</table>
MultiSeq: Display and Edit Metadata

- External databases are cross-referenced to display metadata such as taxonomic information and enzymatic function
- Changes to metadata are preserved for future sessions
- **Electronic Notebook**: Notes and annotations about a specific sequence or structure can be added
Acknowledgements

- Elijah Roberts
- John Eargle
- Ke Chen
- Jonathan Lai
- Kirby Vandivort
- John Stone

NIH Center for Macromolecular Modeling and Bioinformatics