Introduction to evolutionary concepts and VMD/MultiSeq - Part I

Characterizing your systems

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

NIH Resource Macromolecular Modeling and Bioinformatics
Atlanta Workshop 2011

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.”

Evolutionary profiles for protein structure & function prediction

Signatures ribosomal evolution

UPT - Woese 16S rRNA

LSU (23S rRNA + rproteins)
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**


Aquaporin Superfamily: Bacterial & Eucaryal

- **GLP cluster**
- **Glycerol transport**
- **Water transport**

Heymann and Engel *News Physiol. Sci.* (1999)
Archaeal AqpM *M. Marburgensis*, IBC 2003, PNAS 2005

<table>
<thead>
<tr>
<th>Aquaporin</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQP0 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP1 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP2 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP3 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP4 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP5 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP6 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP7 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP8 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>AQP9 HUMAN</td>
<td>483 Amino Acids</td>
</tr>
<tr>
<td>GLP</td>
<td>483 Amino Acids</td>
</tr>
</tbody>
</table>

Ruler: 180, 200, 210, 220, 240, 260, 280, 300
Protein (RNA) Folding, Structure, & Function

Protein:RNA Complexes in Translation
Evolution, Structure, and Dynamics

“Evolution SepRS/CysRS”, PNAS 2005
“Dynamic Signaling Network”, PNAS 2009
“Exit Strategy Charged tRNA” JMB 2010

“Dynamical Recognition Novel Amino Acids”, JMB 2008
“Signatures ribosomal evolution” PNAS 2008, BMC 2009
“Whole cell simulations on GPUs” IEEE 2009
“Dynamics of tRNA” FEBS 2010
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

*G* scattered around gaps

ClustalW alignment
~ 5 minutes

MAFFT* alignment
~ 30 seconds
More sequences!

* Maft* Katoh, Misawa, Kuma,Miyata, NAR 2002, 2005
STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs
   • Multiple Sequence alignment
   • Ridged Body “Scan”

2. Refine Initial Alignment & Produce Multiple Structural Alignment

   \[ P_i^j = \left( e^{-\frac{d_{ij}}{2E_i}} \right) \left( e^{-\frac{s_{ij}}{2E_i}} \right) \]

   - probability that residue i on structure A is equivalent to residue j on structure B.
   - \( d_{ij} \) = distance between i & j
   - \( s_{ij} \) = conformational similarity, function of rms between i-1, i, i+1 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.


Multiple Structural Alignments

STAMP – cont’d

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

\[ S_C = \frac{S_p L_p - l_A L_p - l_B}{L_p L_A L_B} \]
\[ S_p = \sum_{alt.\ path} P_i^j \]
\[ L_{p,i}, L_{A,B}, l_{A,B} = \text{length of alignment, sequence A, sequence B} \]
\[ l_{A,B} = \text{length of gaps in A and B.} \]

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.
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

Look for horizontal gene transfer events

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

Woese, Olsen, Ibba, Soll MMBR 2000

**Protein Structure Similarity Measure**

\[ Q_H \text{ Structural Homology} \]

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

\[ Q_H = \frac{N}{2} [q_{aln} + q_{gap}] \]

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

Structural Similarity Measure: The effect of insertions

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

\[
q_{\text{gap}} = \sum_{\gamma_j} \sum_{\delta_j} \max \left\{ \exp \left[ -\frac{(r_{\gamma,j} - r_{\delta,j})^2}{2\sigma^2} \right], \exp \left[ -\frac{(r_{\gamma,j} - r_{\delta,j})^2}{2\sigma^2} \right] \right\}
\]

Structure encodes evolutionary information!

sequence-based phylogeny

structure-based phylogeny

20 changes

archaeal helix extensions, insertion

Da - AspRS archaeal genre

Db - AspRS bacterial genre

JMB 2005

MMBR 2003, 2000
Structure reveals distant evolutionary events

Class I AARSs
Class II AARSs

Sequences define more recent evolutionary event

Class I Lysyl-tRNA Synthetase
Class II Lysyl-tRNA Synthetase

Conformational changes in the same protein.

Structures for two different species.

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

Q_H = 0.80
Sequence identity = 1.00

ProRS
M. jannaschii, 2.55 Å.
M. thermoautotrophicus, 3.20 Å.

Q_H = 0.89
Sequence identity = 0.69
### Relationship Between Sequence & Structure

- **sequence identity > 20%**
  - 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

- Multidimensional QR factorization of alignment matrix, $A$.

- Economy of information
  - 16 representatives

**Multidimensional QR** factorization of alignment matrix, $A$.

$A = \begin{bmatrix} f_{i_{als}} & f_{i_{rels}} & \cdots & f_{i_{proteins}} \end{bmatrix}$

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: $(x_{C_n}, y_{C_n}, z_{C_n}, 0)$
- Gapped position: $(0, 0, 0, g)$

**Gap Scaling**
$$g = \gamma \left( \sum ||F_i||^2 + \sum ||Z||^2 \right)$$

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

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

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

- $A = (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- $B = (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- $C = (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$
- GAP = $(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)$

Alignment is a Matrix with Linearly Dependent Columns

$$A = \begin{bmatrix} d=1 \\ d=2 \\ d=3 \\ d=N \end{bmatrix}$$

$$Q_{m \times n} A_{n \times N} P = \tilde{R}_{d}$$

Class I AARSs

 evolutionary events

- **5 Subclasses**
- **Specificity** – 11 Amino acids
- **Domain of life** A,B,E

![Diagram showing the relationship between Subclass, Specificity, and Domain of life for Class I AARSs](image-url)
Profile of the ILMV Subclass

How many sequences are needed to represent the Subclass ILMV?

If each of ILMV were 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
Design - Evolution of Structure and Function in Class II

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,..)

Choose MAFFT to perform multiple sequence alignment
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 /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

Variation in structures
Variation in sequences
# 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.

Secondary structure prediction

- Integration with PSIPRED* to predict secondary structure of sequences.
- Compare to VMD STRIDE predictions from structures.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>350</th>
<th>360</th>
<th>370</th>
<th>380</th>
</tr>
</thead>
<tbody>
<tr>
<td>HyFlex, S4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermus, S4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecol, S4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Modeling of Helicobacter pylori ribosomal protein S4 using two known bacterial structures from Thermus thermophilus and Escherichia coli.

Zinc-binding site replaced by salt bridge in H. pylori.

PSIPRED installation

- PSIPRED is not included with VMD, must be installed locally.
- Configured in the MultiSeq software preferences dialog (File->Preferences).

Requires a sequence database filtered for problematic regions. Here using Swiss-Prot for relatively fast predictions.
Export Modeller compatible alignments

- MultiSeq can automatically export SIF alignment files compatible with Modeller.

> P1; Hpylori_S4
sequence: Hpylori_S4:0000:0.00
MARYRGAVERLERRFGVSLALKGE-RRLSGKSALDKRAYGPGQHGQR-RAKTSDYGLQLK
EKQKAKMMYGISEKQFRSIFVEANRLDGNTGENLIRLLERRLVVVFYRMGFGATRAEQ
LVSHKAVLSFQYERFGQEQYRSQTVLEAGD

> P1; Thermus_S4
structureX: Thermus_S4:209:0.00
-GRYIGPVCRLCRREGVKLYLKGE-RCYSPKCAMERRPYPPGQHGQKRARRPSDYAVRLR
EKQKLRIYRIQFQRMNLTEAEZKQVTVGFPLLLESRLDNYVVRFLGAFVERQRAEQ
LVAHGRGRRVQVFQBPQDIEVNVSLGOKFMLEHAEHKGKRVQVPLGSLGO

> P1; Ecoli_S4
structureX: Ecoli_S4:120:0.00
ARYLGPKLKLSRREGTDLFLKSGVRAIDTKCKIE---QAPGQHGAR-KPRLSDYGVQLR
EKQKVRRIYGVLERQFRNYYKEAARLKGNTGENLLALLEGRLDNVVYRMGFGATRAEQ
LVSHKAIWGVVIASQYVPSRQVHSQRYKAEALAEQKRRKTLVELEDOAG

```
a = mymodel(env, alnfile='alignment.ali', knowns=('Ecoli_S4', 'Thermus_S4'), sequence='Hpylori_S4')
a.starting_model = 1
a.ending_model = 20
a.make()
```

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.

Manipulate branches to simplify the tree

- Manually collapse by node.
- Automatically collapse clades that are alike according to taxonomy, enzyme, temperature class, and/or MultiSeq grouping.
- Set the root of the tree manually, if known from external sources.

Combined phylogenetic tree and genome content analysis of ribosomal protein S4 for all complete bacterial genomes.

Roberts, Chen, ZLS,
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.

Location | Strand | Length | PID | Gene | Synonym | Code | COG | Product
--- | --- | --- | --- | --- | --- | --- | --- | ---
3437638..3438021 | - | 127 | 16131173 | rplQ | b3294 | - | COG2032 | 50S ribosomal subunit protein L17
3438062..3439051 | - | 329 | 16131174 | rpoA | b3295 | - | COG2022 | RNA polymerase, alpha subunit
3439071..3439687 | - | 206 | 16131175 | rpsD | b3296 | - | COG0522 | 30S ribosomal subunit protein S4
3439121..3440120 | - | 129 | 16131176 | rpsK | b3297 | - | COG0105 | 30S ribosomal subunit protein S1
3440137..3440693 | - | 118 | 16131177 | rpsM | b3298 | - | COG0099 | 30S ribosomal subunit protein S5
3440520..3440756 | - | 36 | 16131178 | rpmJ | b3299 | - | COG0257 | 50S ribosomal subunit protein L3
3440578..3442119 | - | 443 | 16131179 | secY | b3300 | - | COG0201 | preprotein translocase membrane subunit
3442127..3442956 | - | 144 | 16131180 | rplO | b3301 | - | COG0200 | 50S ribosomal subunit protein L15
3442565..3442744 | - | 59 | 16131181 | rpsE | b3302 | - | COG0087 | 30S ribosomal subunit protein S5
3442748..3443251 | - | 167 | 16131182 | rpmD | b3303 | - | COG0098 | 30S ribosomal subunit protein L30

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.

```bash
proc draw_genome_context_of_phylogeny {args} {
    # Load the sequences.
    set alignment [:SeqData::Fasta::loadSequences $alignmentFilename]
    # Load the tree.
    set tree [:PhylTree::Newick::loadTreeFile $treeFilename]
    # Reorder the alignment by the tree.
    set treeAlignment {}
    set leafNodes [:PhylTree::Data::getLeafNodes $tree]
    foreach node $leafNodes {
        set foundNode 0
        set nodeName [:PhylTree::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

```python
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
    }
}
```

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.
BLAST DB Searching

- Import sequence data directly from BLAST databases
- Search using a single sequence or an EP profile
- Filter results based on taxonomy or redundancy (QR)

Protein sequence alignment

How do I align two similar, but different sequences?

Sequence 1: $a_1 a_2 a_3 \cdots a_4 a_5 \cdots a_n$
Sequence 2: $c_1 - c_2 c_3 c_4 c_5 - \cdots c_m$

There exist fast web tools, e.g., BLAST search: http://www.ncbi.nlm.nih.gov/
See also Blastn, Psi-Blast, ...
Sequences from Swiss-Prot, NCBI, JGI, ....

Structures from PDB, CATH, SCOP, ....

### NiceProt View of Swiss-Prot:
P47865

<table>
<thead>
<tr>
<th>Entry name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQPI_BOVIN</td>
<td>Aquaporin-CHIP</td>
</tr>
</tbody>
</table>

#### Synonyms
- Water channel protein for red blood cells and kidney proximal tubule
- Aquaporin 1
- Water channel protein CHIP20

#### Gene name
- Name: AQPI

#### Taxonomy
- Bos taurus (Bovine) [TaxID: 9823]

#### References
- [1] SIQUES FROM NUCLEIC ACID
- TBD | End of character summary |

---

**Final Blast Result: Sequence Alignment**

```
>qi|451595501|sp|08817|A005_P6EPK    O Aquaporin 2
Length = 230
Score = 119 bits (29%), Expect = 6e-27
Identities = 70/186 (37%), Positives = 105/186 (56%), Gaps = 12/186 (6%)

Query:  53  VSLAFGLGIAATLQAQQGVSGLAGMNPAVTGLLILSCQSELHAIHY1AQCVGAIVATAT 112
        W+ ASGL++ T+A ++ GHSIS RLPAV+ GL++ + + Y+IAQ +GAI+A +
        Sbjct:  40  VPAFAGLVTLMFAFAGHISCLCHNPASPLGLVGGFRPAKELLFVIAQCGAIAAGV 99

Query:  113  LSGITSLIP--DNSLGLL---MALAP---GVMGCQLGLIIGTQQLAVLCLATGDBRRD 164
        + I S + S GL N A G G G E++ T ++ ++ TD R
        Sbjct:  100  IYLLASQKAGFSEALSLAGSNGYADHSPGGYTLGAFVSFVVTAMKFLVVMAGTARAP-- 158

Query:  165  LGSSCPPLAIGFSVALCHLLAIDVTGQGGINPARRSFSGSSVIPNHP--QNFWIFWUGPPLCGA 222
        G R+AIQ ++ L HL++I T ++PARS G ++ + Q W+PMW + IGA
        Sbjct:  159  --AGFAAIRAGAALTLHLSIPFVMNTSVPNARSTGPALFVGQWLQQLFLFVMAYFIAAGA 217

Query:  223  LAVLII 228
        + +
        Sbjct:  218  IGAYAL 223
```

Search returns approximate alignments - needing refinement!
- Clustal, Muscle, MAFT, Tcoffee, pileup, Smith-Waterman, and manual editing in sequence editor
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

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
• Kirby Vandivort
• John Stone
• Michael Bach
• NIH Resource for Macromolecular Modeling and Bioinformatics