# Tools and Algorithms in Bioinformatics

GCBA815, Fall 2013

## *Week4: Multiple Sequence Alignment*

Babu Guda, Ph.D.

Department of Genetics, Cell Biology & Anatomy

Bioinformatics and Systems Biology Core

University of Nebraska Medical Center



## Terminology

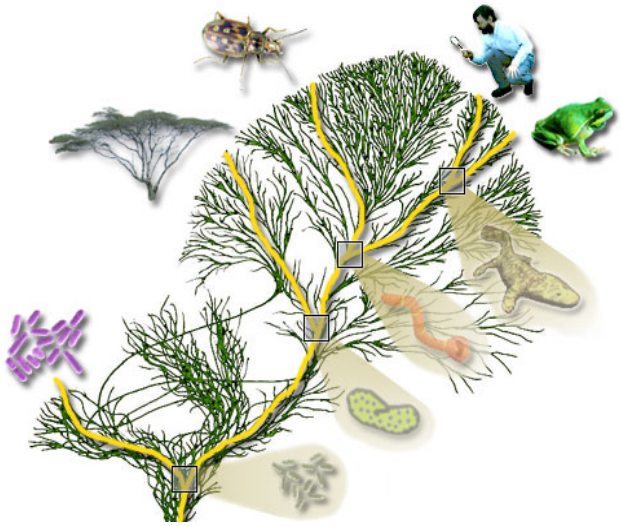
- Homologous : Similar
- Paralogous : Present in the same species, diverged after gene duplication.
- Orthologous: Present in different species, diverged after speciation.
- Xenologous: Genes acquired by horizontal gene transfer
- Analogous: Similarity observed by convergent evolution, but not by common evolutionary origin.

## Evolution of Genomes and Genetic Variation

- Variation between species
  - Recombination, Cis-regulatory elements
  - Point mutations/ insertions/deletions
  - Evolutionary selection → speciation
- Variation within species:
  - Phenotype = Genotype + Environment
  - SNPs, haplotypes, aneuploidy, etc.
- Variation at cellular level:
  - Spatial state (Tissue/Subcellular location)
  - Temporal state (Stages in life cycle)
  - Physiological state (normal/disease)
  - External stimuli



## Phylogenetic Tree of Life

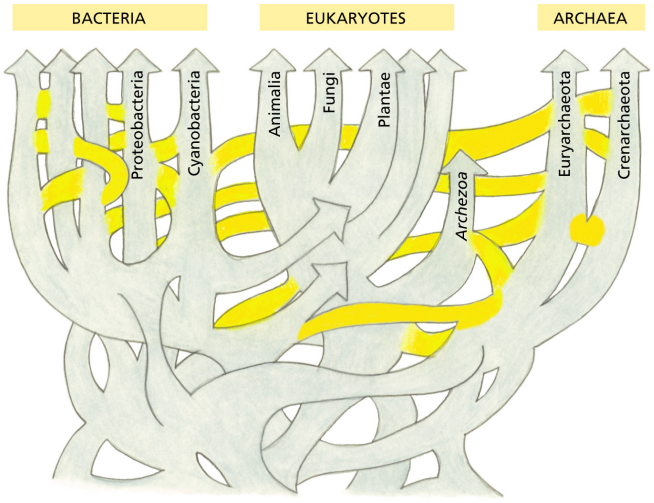


## Multiple Sequence Alignments

- Aligning homologous residues among a set of sequences, together in columns
- Multiple alignments exhibit structural and evolutionary information among closely related species i.e., families
    - A column of aligned residues in a conserved region is likely to occupy similar 3-D positions in protein structure
- Patterns or motifs common to a set of sequences may only be apparent from multiple alignments
- Necessary for building family profiles, phylogenetic trees and extracting evolutionary relationships
- Useful for homology modeling if at least one member of the family has a known 3-D structure



## Phylogenetic Complexity



## Multiple alignments from different VGC protein families

### Sodium VGC

```
YNMTAEFEEMLQVGNLVFTGIFTAEMTFKIIALD
HPMTPQFEHVLAVGNLVFTGIFTAEMFLKLIAMD
FPMSEELRSLLHVGNLVFIGIYTIELILKIIAMH
YPMTEHFDNVLSVGNLVFTGIFTAEMVLKLIAMD
YPMTEEFNNVLNVGNLVFTGIFTAEMVLKLIALD
YPMTEEFDYMLSVGNLVFTGIFAAEMFFKLIAMD
DEQSAEKTKILNKINQFFVAVFTGECVMKMFALR
DDQSEEKTKILGKINQFFVAVFTGECVMKMFALR
DNQSEEKTKVLGRINQFFVAVFTGECVMKMFALR
DEQGEEKTKVLGRINQFFVAVFTGECVMKMFALR
YNQPKAMKSILDHLNWVFVVIFTLECLIKIFALR
EGQPNEVKKIFDILNIVFVVIFTVECLIKVFALR
ADQPKDVKKTFDILNIAFVVIFTIECLIKVFALR
DDQSEYVTTILSRINLVFIVLFTGECVLKLISLR
```

### Calcium VGC

```
HYYFTNSWNIFDFVVVILSIVGTVLSDIIQ-
HYYFTVGWNIFDFVVVIISIVGMFLADLIE-
HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
HYYFTVGWNIFDFVVVILSIVGMFLAEMIE-
HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
LEYFKYGMNVFDFVIVVFSIAVIIMIEYDE-
WFYFKDPWNVFDFSVVVFSTIAWILQFFES-
FYYFKNPWNVFDFIVVILSVVGSTMNEVIK-
WYYFKEPWNVFDCVVVTLSILGIAIKDLIA-
```

### Potassium VGC

```
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
VQVFRIMRILRVLKLARHSTGLQS
LKVVRLLRLGRVVRKLDRYLEYGA
FVTLRVFRVFRIFKFSRHSQGLRI
LEFFSIIRIMRLFKLTRHSSGLKI
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
```



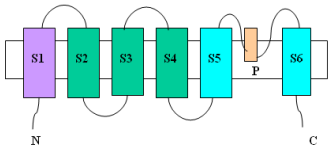
Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 352 (2007) 292–298
BBRC

**Conserved motifs in voltage-sensing and pore-forming mo of voltage-gated ion channel proteins**

Purnima Guda [a], Philip E. Bourne [b], Chittibabu Guda [a,*]

[a] GenNYis Center for Excellence in Cancer Genomics and Department of Epidemiology and Biostatistics, State University of New York at Albany, One Discovery Drive, Rensselaer, NY 12144-3456, USA
[b] Department of Pharmacology, University of California San Diego, La Jolla, CA 92093, USA

Received 13 October 2006
Available online 13 November 2006




- Large scale data analysis using profile-profile alignments to detect patterns in the VGC family of proteins

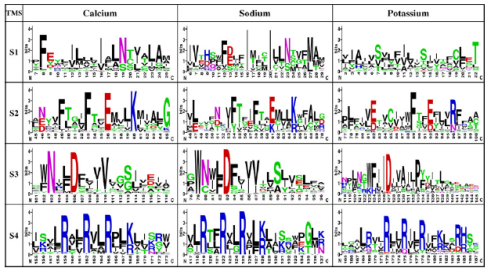



Fig. 1. Conserved motifs in the voltage-sensing module of calcium, sodium, and potassium ion channel proteins.

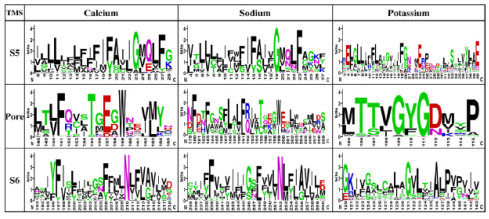



Fig. 2. Conserved motifs in the pore-forming module of calcium, sodium, and potassium ion channel proteins.

## Scoring a Multiple Alignment

- Some positions are more conserved than others
- Sequences are not independent, but are related by a phylogeny

```
   430667            100.0%    MTDREQFFRDVNKIVIKIGTSSITRKGCDHTKENCNIDPAFMESIAAQVY
 1 SWALL:PROB_MEIRU   38.4%    -------------LPLTAQTHRRLVVKVGSAVLSGPQGRQHQLAIAAQVA
 2 SWALL:Q9KPT8       38.4%    MTTNQRSIKAPQTVVVKLGTSVLTGG-------TLALDRAHMVELARQCA
 3 SWALL:AAG07953     36.5%    ---MRDKVTGARRVVVKIGSALLTADGR-------GLDRNAMAVVVEQMV
 4 SWALL:Q9RTD8       38.6%    -----------MRVVLKLGTSVLTA-GTDRLHRPRLVD--LMRDIAA---
 5 SWALL:PROB_THETH   34.9%    --------------RVEEGVLIPRPETEGLVELALGLPLPPAPRIARQVA
 6 SWALL:PROB_HAEIN   38.0%    --------MNKKTIVVKFGTSTLTQGS---PKLN---SPHMME-IVRQIA
 7 SWALL:PROB_ECOLI   36.6%    -------MSDSQTLVVKLGTSVLTGG-------SRRLNRAHIVELVRQCA
 8 SWALL:Q9KCR4       37.1%    --------MKRQRIVIKIGSSSLT------TKQGA-LDLEKLEGYVRAIV
 9 SWALL:PROB_SYNY3   36.1%    ----MTMAMQPQTLVIKIGTSSLARP------ETGQLALSTIAALVETVC
10 SWALL:PROB_BACSU   37.2%    --------MKKQRIVVKIGSSSLTnkG--------SIDEAKIREHVQAIS
11 SWALL:PROB_SERMA   35.3%    -------MNGSQTLVVKLGTSVLTGG-------SLRLNRAHIVELVRQCA
12 SWALL:O07509       34.3%    ----MTPDTSMKRVVVKIGSSSLTSL-------HGEISIRKLEALVDQVV
13 SWALL:Q9RDJ9       34.7%    MAGARQAVGEARRIVVKVGSSSLTTAAG-------GLDADRVDALVDVLA
14 SWALL:PROB_MYCTU   33.5%    RSPHRDAIRTARGLVVKVGTTALTTP-------SGMFDAGRLAGLAEAVE
15 SWALL:PROB_AQUAE   35.0%    -----------MRIVFKIGSNLL------ETDEG-DIDLSFLSKLAEGIK
```



**Pair-wise**

```
LVD--LMRDIAA----VSAQGHE
GLPLPPAPRIARQVAALREEGRE
```

**Multiple**

```
KIGTSSITRKGCDHTKENCNIDPAFME
TAQTHRRLVVKVGSAVLSGPQGRQHQL
KLGTSVLTGG-------TLALDRAHMV
KIGSALLTADGR-------GLDRNAMA
KLGTSVLTA-GTDRLHRPRLVD--LMR
```

- Most of the multiple alignment algorithms assume that the individual columns of an alignment are statistically independent

Total Alignment Score $S = \Sigma_i S(m_i) + G$

where, $m_i$ is column $i$ of the multiple alignment $m$,

$S(m_i)$ is the score for column $i$,

$G$ is gap penalty

Table 1 - The log odds matrix for 250 PAMs (multiplied by 10)

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | -2 | 0 | 0 | -4 | 1 | -1 | -1 | -1 | -2 | -1 | 0 | 1 | 0 | -2 | 1 | 1 | 0 | -6 | -3 |
| C | | 12 | -5 | -5 | -4 | -3 | -3 | -2 | -5 | -6 | -5 | -4 | -3 | -5 | -4 | 0 | -2 | -2 | -8 | 0 |
| D | | | 4 | 3 | -6 | 1 | 1 | -2 | 0 | -4 | -3 | 2 | -1 | 2 | -1 | 0 | 0 | -2 | -7 | -4 |
| E | | | | 4 | -5 | 0 | 1 | -2 | 0 | -3 | -2 | 1 | -1 | 2 | -1 | 0 | 0 | -2 | -7 | -4 |
| F | | | | | 9 | -5 | -2 | 1 | -5 | 2 | 0 | -4 | -5 | -5 | -4 | -3 | -3 | -1 | 0 | 7 |
| G | | | | | | 5 | -2 | -3 | -2 | -4 | -3 | 0 | -1 | -1 | -3 | 1 | 0 | -1 | -7 | -5 |
| H | | | | | | | 6 | -2 | 0 | -2 | -2 | 2 | 0 | 3 | 2 | -1 | -1 | -2 | -3 | 0 |
| I | | | | | | | | 5 | -2 | 2 | 2 | -2 | -2 | -2 | -2 | -1 | 0 | 4 | -5 | -1 |
| K | | | | | | | | | 5 | -3 | 0 | 1 | -1 | 1 | 3 | 0 | 0 | -2 | -3 | -4 |
| L | | | | | | | | | | 6 | 4 | -3 | -3 | -2 | -3 | -3 | -2 | 2 | -2 | -1 |
| M | | | | | | | | | | | 6 | -2 | -2 | -1 | 0 | -2 | -1 | 2 | -4 | -2 |
| N | | | | | | | | | | | | 2 | -1 | 1 | 0 | 1 | 0 | -2 | -4 | -2 |
| P | | | | | | | | | | | | | 6 | 0 | 0 | 1 | 0 | -1 | -6 | -5 |
| Q | | | | | | | | | | | | | | 4 | 1 | -1 | -1 | -2 | -5 | -4 |
| R | | | | | | | | | | | | | | | 6 | 0 | -1 | -2 | 2 | -4 |
| S | | | | | | | | | | | | | | | | 2 | 1 | -1 | -2 | -3 |
| T | | | | | | | | | | | | | | | | | 3 | 0 | -5 | -3 |
| V | | | | | | | | | | | | | | | | | | 4 | -6 | -2 |
| W | | | | | | | | | | | | | | | | | | | 17 | 0 |
| Y | | | | | | | | | | | | | | | | | | | | 10 |



# Progressive Alignment Methods

- Most commonly used approach for multiple sequence alignment
- Initially, standard pair-wise alignments are done and successive sequences are added to the first alignment until all sequences have been aligned
- The drawback is that alignments are heuristic, global score is not optimized in the algorithm and optimum alignments are not guaranteed
- The advantage is that these are very fast, and in many cases result in reasonable alignments

Several progressive alignment strategies are available. The algorithms differ in several ways such as

- the way the order of the alignment is chosen
- whether the progression involves alignment of sequences to a single growing alignment or whether sub-families are aligned first and these alignments are aligned to other alignments
- the procedure used to score sequences or alignments



## Basic progressive alignment procedure

- Calculate alignment score for all pair-wise combinations using Dynamic Programming
- Determine distances from scores for all pairs of sequences and build a distance matrix
- Use a distance-based method to construct a 'guide-tree'
- Add sequences to the growing alignment using the order given by the tree
- Most of the multiple alignment methods differ in the way the guide tree is constructed

## Feng-Doolittle Method

Get a distance matrix

Calculate pair-wise scores with DP method, convert raw scores into pair-wise distances using the following formula

$$D = -\ln S_{eff}$$

$$S_{eff} = [S_{obs} - S_{rand}] / [S_{max} - S_{rand}]$$

where, -$S_{eff}$ is the effective score

-$S_{obs}$ is the similarity score(DP score) between a pair

-$S_{max}$ is the max score, average of aligning either sequence to itself

-$S_{rand}$ is the background noise, obtained by aligning two random sequences of equal length and composition



**Feng-Doolittle Method …..**

From pair-wise distances, build a matrix for all sequence combinations as follows

| | Human | Chimp | Gorilla | Orang |
|---|---|---|---|---|
| Human | 0 | 88 | 103 | 160 |
| Chimp | | 0 | 106 | 170 |
| Gorilla | | | 0 | 166 |
| Orang | | | | 0 |

• The number of pair-wise distances are *N(N-1)/2*

**Feng-Doolittle Method …..**


**Alignment of sequences**

- The alignment order is determined from the order sequences were added to the guide tree
- First 2 sequences from the node are added first. In this case, C-H are aligned according to the standard DP algorithm
- Next, G is aligned to CH as the best of G(CH) and (CH)G alignments
- Assuming that of the above two, G(CH) has the best score, O is aligned to G(CH) as the best of O(G(CH)) and G(O(CH))
- Again, higher similarity score determines which one is the best. This process is repeated iteratively until all sequences are aligned.
- As you notice, the order of sequences in the output are not the same as you see in the guide tree.



## CLUSTALW - Multiple sequence alignment

http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html

- **The most popular multiple alignment method with many optional parameters**
- **Obtains a distance matrix by using scores from pair-wise DP method**
- **Guide tree is built using 'neighbor joining' method**
- **Sequences are weighed to compensate for biased representation in larger families**
- **Substitution matrices change on the fly as the alignment progresses; closely related sequences are aligned with 'hard' matrices (e. g. BLOSUM80) and distant sequences are aligned with 'soft' matrices (e.g. BLOSUM40)**
- **Position specific gap penalties are used similar to profiles**
- **Guide tree may be adjusted on the fly to defer the alignment of low-scoring sequences**