

Wright State University

## CORE Scholar

---

Computer Science and Engineering Faculty  
Publications

Computer Science & Engineering

---

2003

### Protein Alignment Scoring - PAM and BLOSUM

Dan E. Krane

*Wright State University - Main Campus*, dan.krane@wright.edu

Michael L. Raymer

*Wright State University - Main Campus*, michael.raymer@wright.edu

Follow this and additional works at: <https://corescholar.libraries.wright.edu/cse>



Part of the [Computer Sciences Commons](#), and the [Engineering Commons](#)

---

#### Repository Citation

Krane, D. E., & Raymer, M. L. (2003). Protein Alignment Scoring - PAM and BLOSUM. .  
<https://corescholar.libraries.wright.edu/cse/388>

This Presentation is brought to you for free and open access by Wright State University's CORE Scholar. It has been accepted for inclusion in Computer Science and Engineering Faculty Publications by an authorized administrator of CORE Scholar. For more information, please contact [library-corescholar@wright.edu](mailto:library-corescholar@wright.edu).

# Sequence Alignments Revisited

---

- Scoring nucleotide sequence alignments was easier
  - Match score
  - Possibly different scores for transitions and transversions
- For amino acids, there are many more possible substitutions
- How do we score which substitutions are highly penalized and which are moderately penalized?
  - Physical and chemical characteristics
  - Empirical methods

# Scoring Mismatches

---

- Physical and chemical characteristics
  - $V \rightarrow I$  – Both small, both hydrophobic, conservative substitution, small penalty
  - $V \rightarrow K$  – Small  $\rightarrow$  large, hydrophobic  $\rightarrow$  charged, large penalty
  - *Requires some expert knowledge and judgement*
- Empirical methods
  - How often does the substitution  $V \rightarrow I$  occur in proteins that are known to be related?
    - Scoring matrices: PAM and BLOSUM

# PAM matrices

---

- PAM = “Point Accepted Mutation” interested only in mutations that have been “accepted” by natural selection
- Starts with a multiple sequence alignment of very similar (>85% identity) proteins. Assumed to be homologous
- Compute the *relative mutability*,  $m_i$ , of each amino acid
  - e.g.  $m_A$  = how many times was alanine substituted with anything else?

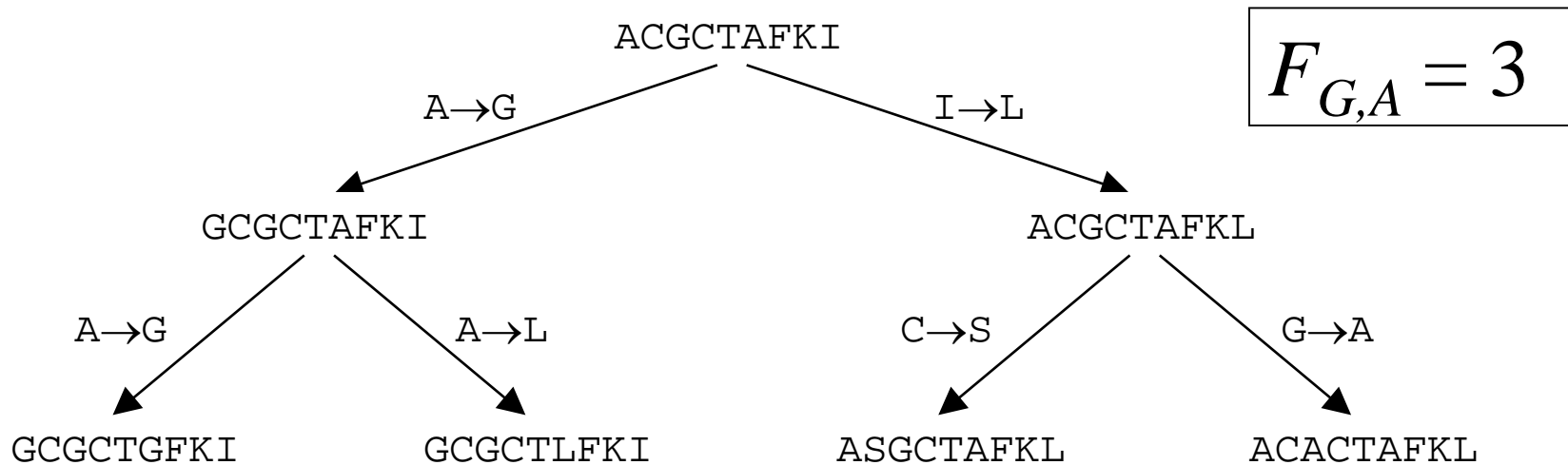
# Relative mutability

---

- ACGCTAFKI  
GCGCTAFKI  
ACGCTAFKL  
GCGCTGFKI  
GCGCTLFKI  
ASGCTAFKL  
ACACTAFKL
- Across *all pairs* of sequences, there are 28  
A → X substitutions
- There are 10 ALA residues, so  $m_A = 2.8$

# Pam Matrices, cont'd

- Construct a phylogenetic tree for the sequences in the alignment

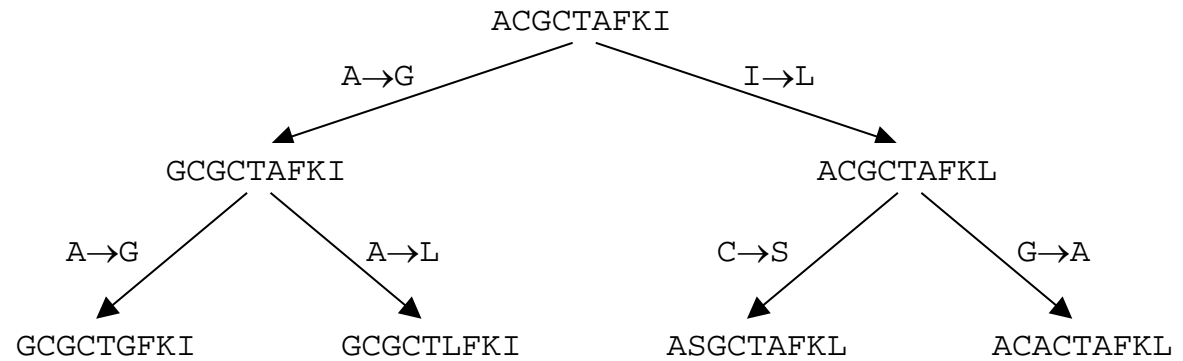


- Calculate substitution frequencies  $F_{X,X}$
- Substitutions may have occurred either way, so  $A \rightarrow G$  also counts as  $G \rightarrow A$ .

# Mutation Probabilities

- $M_{i,j}$  represents the probability of  $J \rightarrow I$  substitution.

$$M_{ij} = \frac{m_j F_{ij}}{\sum_i F_{ij}}$$



- $M_{G,A} = \frac{2.7 \times 3}{4} = 2.025$

# The PAM matrix

---

- The entries,  $R_{i,j}$  are the  $M_{i,j}$  values divided by the frequency of occurrence,  $f_i$ , of residue  $i$ .
- $f_G = 10 \text{ GLY} / 63 \text{ residues} = 0.1587$
- $R_{G,A} = \log(2.025/0.1587) = \log(12.760) = 1.106$
- The log is taken so that we can add, rather than multiply entries to get compound probabilities.
- *Log-odds* matrix
- Diagonal entries are  $1 - m_j$



# Interpretation of PAM matrices

---

- PAM-1 – one substitution per 100 residues (a PAM unit of time)
- Multiply them together to get PAM-100, etc.
- “Suppose I start with a given polypeptide sequence  $M$  at time  $t$ , and observe the evolutionary changes in the sequence until 1% of all amino acid residues have undergone substitutions at time  $t+n$ . Let the new sequence at time  $t+n$  be called  $M'$ . What is the probability that a residue of type  $j$  in  $M$  will be replaced by  $i$  in  $M'$ ?”

# PAM matrix considerations

---

- If  $M_{i,j}$  is very small, we may not have a large enough sample to estimate the real probability. When we multiply the PAM matrices many times, the error is magnified.
- PAM-1 – similar sequences, PAM-1000 very dissimilar sequences

# BLOSUM matrix

---

- Starts by clustering proteins by similarity
- Avoids problems with small probabilities by using averages over clusters
- Numbering works opposite
  - BLOSUM-62 is appropriate for sequences of about 62% identity, while BLOSUM-80 is appropriate for **more** similar sequences.