University of California, Berkeley- Statistics & Genomics Seminar - 2002-10-03

Correlated Amino Acid Substitutions and Sequence Alignment

Gavin E. Crooks,

Richard Green & Steven E. Brenner

Department of Plant and Microbial Biology
University of California, Berkeley

Funding

NIH

Sloan/DOE Postdoctoral Fellowship in Computational Molecular Biology

Pairwise Sequence Alignment

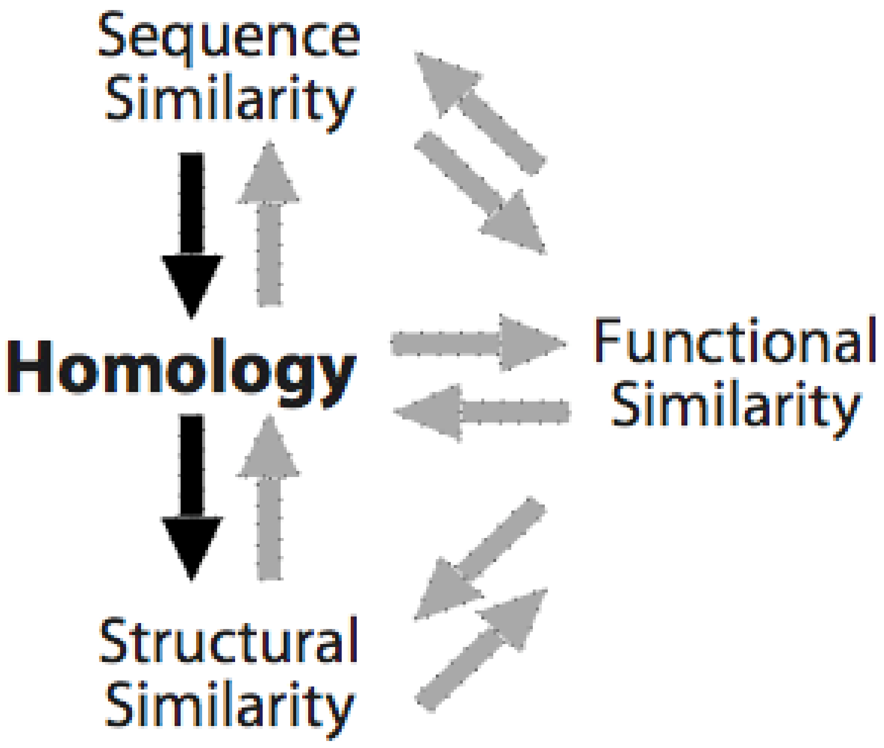
P01958 (Horse Hemoglobin-alpha) vs. P02062 (Horse Hemoglobin-beta)

```
        10        20        30        40         50
 VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLSH-----GSA
  ::. .:. : : :.::. .  : :.::: :... .: :. .:  : :::.     :.
 QLSGEEKAAVLALWDKVNEE--EVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNP
        10        20          30        40        50

     60        70        80        90       100       110
QVKAHGKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHL
.:::::::: ...  .: :::.: :... ::.:: .::.::: ::.::.. :. .:: :.
KVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHF
60        70        80        90       100       110
```

Inferences From Sequence Similarity

- Pairwise Alignment Detects □12% Remote Homologies



Sequence Similarity

Homology

Functional Similarity

Structural Similarity

Alignment Score

Sequence X	AGCACDC-A
Sequence Y	A-CACECTA

$$\text{Score} = g + \sum_{i=1}^{n} s(x_i, y_i)$$

Gap Penalties

Substitution Score

Substitution Matrix

	A	R	N	D	C	Q	E	G	H
A	4	-1	-2	-2	0	-1	-1	0	-2
R	-1	5	0	-2	-3	1	0	-2	0
N	-2	0	6	1	-3	0	0	0	1
D	-2	-2	1	6	-3	0	2	-1	-1
C	0	-3	-3	-3	9	-3	-4	-3	-3
Q	-1	1	0	0	-3	5	2	-2	0
E	-1	0	0	2	-4	2	5	-2	0
G	0	-2	0	-1	-3	-2	-2	6	-2
H	-2	0	1	-1	-3	0	0	-2	8

$$s(x_i, y_i) = c \log \frac{P^s(x_i, y_i)}{q(x_i)\, q(y_i)}$$

Pairwise Alignment Approximations

- Substitution Probabilities are Homogeneous
  - Independent of position
  - Independent of protein, protein family, structure or organism.
- Substitutions are Uncorrelated
  - Independent of surrounding sequence
  - Independent of other substitutions

Substitution Correlations & Sequence Alignment

- Alignment Score
- Efficient Alignment Algorithm
- Doublet Substitution Matrix
- Evaluation

Prior Art

- **Analysis of amino-acid substitution during divergent evolution--The 400 by 400 dipeptide substitution matrix.**

  Gonnet GH, Cohen MA, Benner SA. 1994. *Biochem*. *Biophys*. *Res*. *Comm*. *199*:489-496.

  ✓ Significant correlations in dipeptide substitutions

  × Not enough data to generate a complete substitution matrix

- **Use of residue pairs in protein sequence-sequence and sequence-structure alignments**,
  Jung JS, Lee BK, Protein Science 9 (8): 1576-1588 AUG 2000

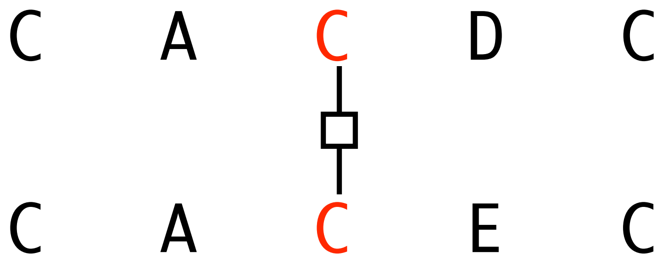
  ✓ homolog detection

  × Cannot find the optimal alignment fast

  × Evaluated 137 sequences

Alignment Score

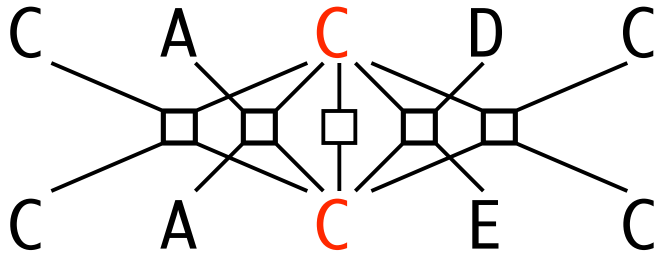
Independent



C A C D C

C A C E C

Correlations



C A C D C

C A C E C

$$\text{Score} = \sum_{i=1}^{n} \ln \frac{P^{s}(x_i,y_i)}{q(x_i)\,q(y_i)} + \sum_{i=1}^{n}\sum_{l=1}^{L} \ln \frac{P_l^{d}(x_i x_{i+l}, y_i y_{i+l})}{q(x_i x_{i+l})\,q(y_i y_{i+l})}\,\frac{q(x_i)\,q(x_{i+l})\,q(y_i)\,q(y_{i+l})}{P^{s}(x_i,y_i)\,P^{s}(x_{i+l},y_{i+l})}$$

Alignment Score With Correlations

$$\text{Score} = g + \sum_{i=1}^{n} s(x_i, y_i) + \sum_{i=1}^{n} \sum_{l=1}^{L} d_l(x_i, x_{i+l}, y_i, y_{i+l})$$

Gap Penalties

Singlet Score

Doublet Score

Restrict doublet distance, e.g. L=3

Doublet BLOSUM Substitution Matrix

- 400x400 doublet substitution matrix
  - 160,000 entries
- BLOCKS 13+ database
  - ≈10^7 independent substitutions

	1	2	3	4	5
CH HA	5	2	0	-1	1
CH HR	3	-3	-2	5	-3
CH HN	5	2	-5	2	3
CH HD	6	3	1	1	-4
CH HC	5	8	8	4	3
CH HQ	3	-1	1	-1	-3
CH HE	5	3	3	0	-1
CH HG	4	7	2	2	1

Data Smoothing: Pseudocounts

counts pseudocounts

$$P_l^d\left(x_i x_{i+l}, y_i y_{i+l}\right) = \frac{n + \square}{N + A} \qquad N = \square\, n \quad A = \square\, \square$$

$$\square = A\, P_l^s\left(x_i, y_i\right) P_l^s\left(x_{i+l}, y_{i+l}\right)$$

- Maximum likelihood estimate of A
  - Multinomial sampling
  - Dirichlet prior
  - Negative hypergeometric data

Finding The Optimal Alignment

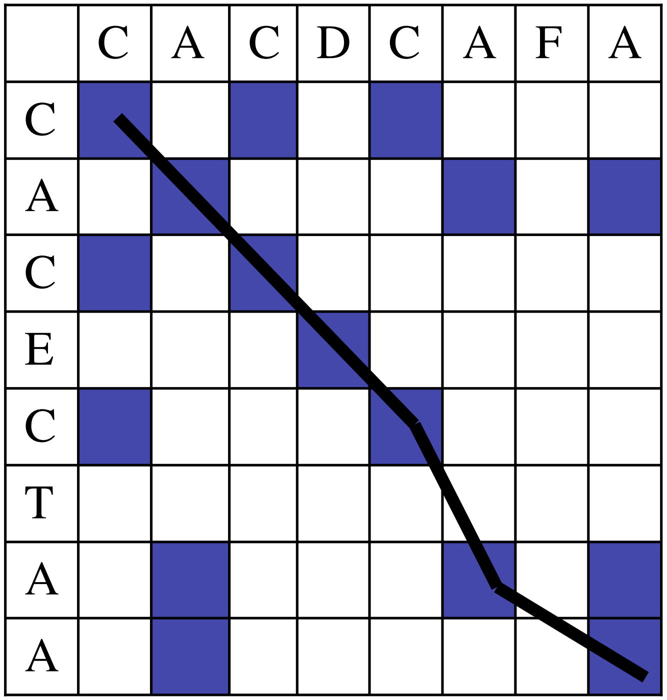
- Dynamic Programming
  - Time ~ O(NM)
  - Memory ~ O(N+M)

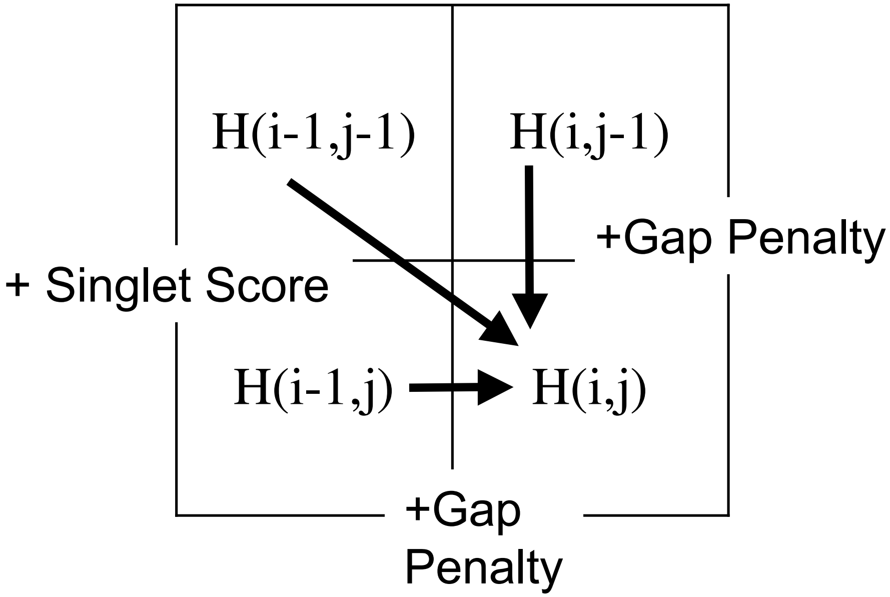
    - N, M : Sequence Lengths

Dynamic Programming

```
CACDC-AFA
CACECTA-A
```

Time ~ O(NM)





H(i-1,j-1)

H(i,j-1)

+Gap Penalty

+ Singlet Score

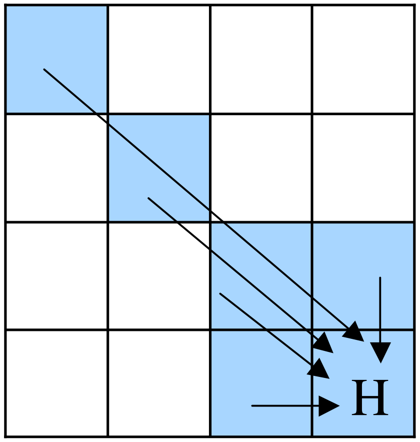
H(i-1,j)

H(i,j)

+Gap Penalty

Finding The Optimal Alignment

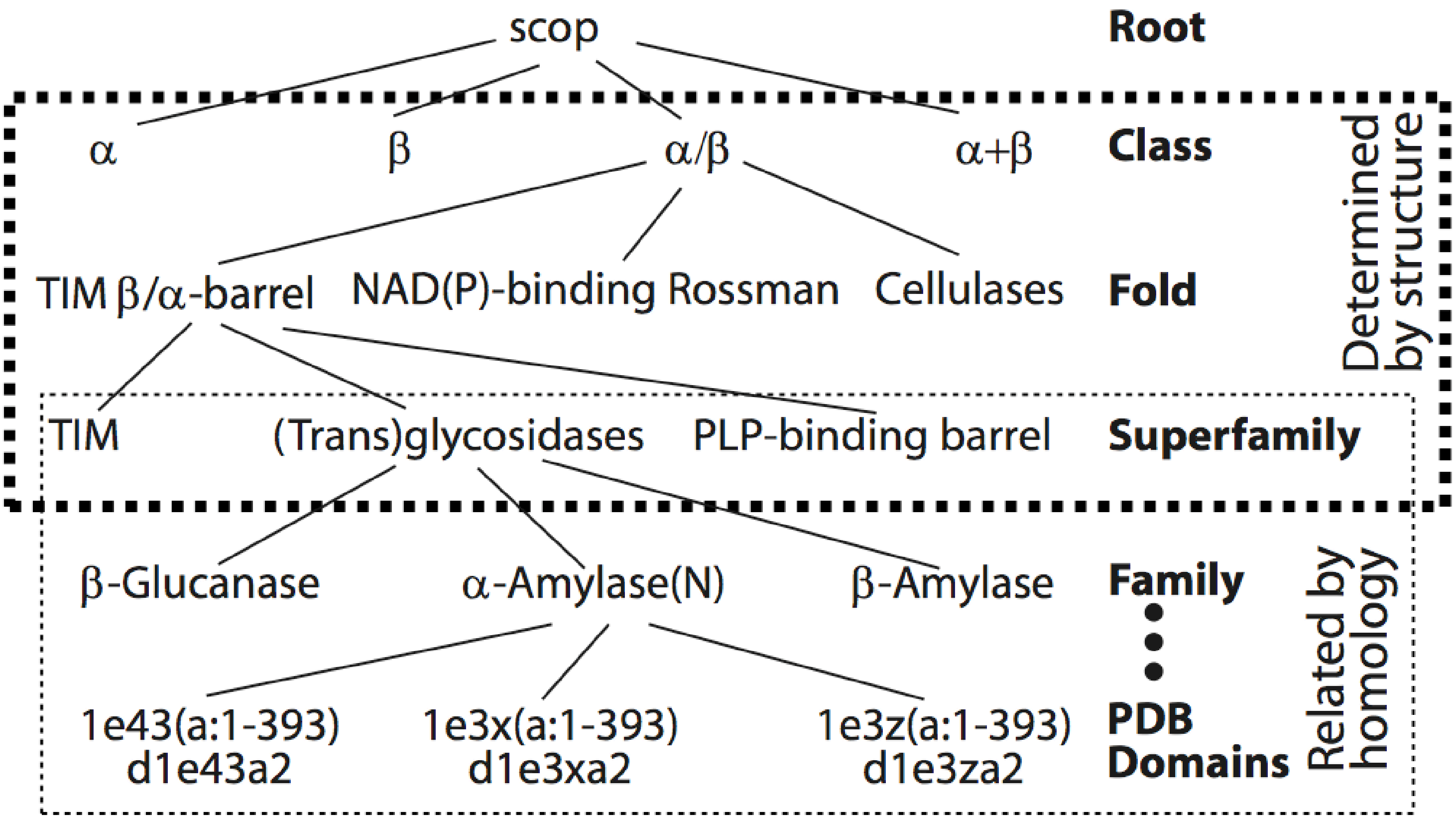
- Dynamic Programming with Short Range Correlations



  - Time ~ O(NML)
  - Memory ~ O((N+M)L)
    - N, M : Sequence Lengths
    - L : Maximum Correlation Length

Evaluation: Remote Homology Detection

- SCOP: Structural Classification of Proteins



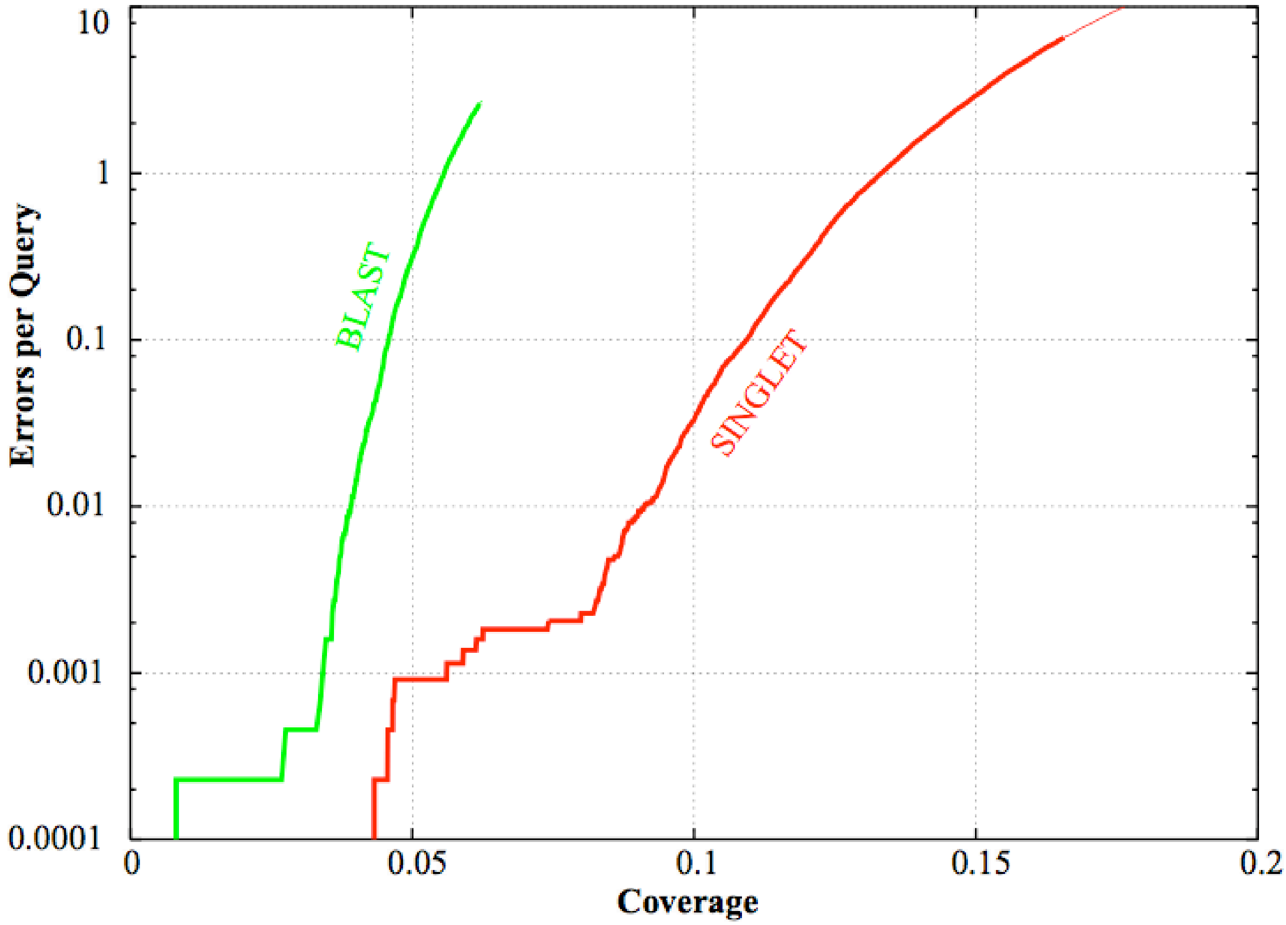
SCOP: Structural Classification of Proteins

- ASTRAL SCOP 1.59 filtered at 40% identity
  - 4380 Protein Domains
  - 1070 Superfamilies
  - 19,184,400 Pairwise Alignments
  - 83,668 Homologies

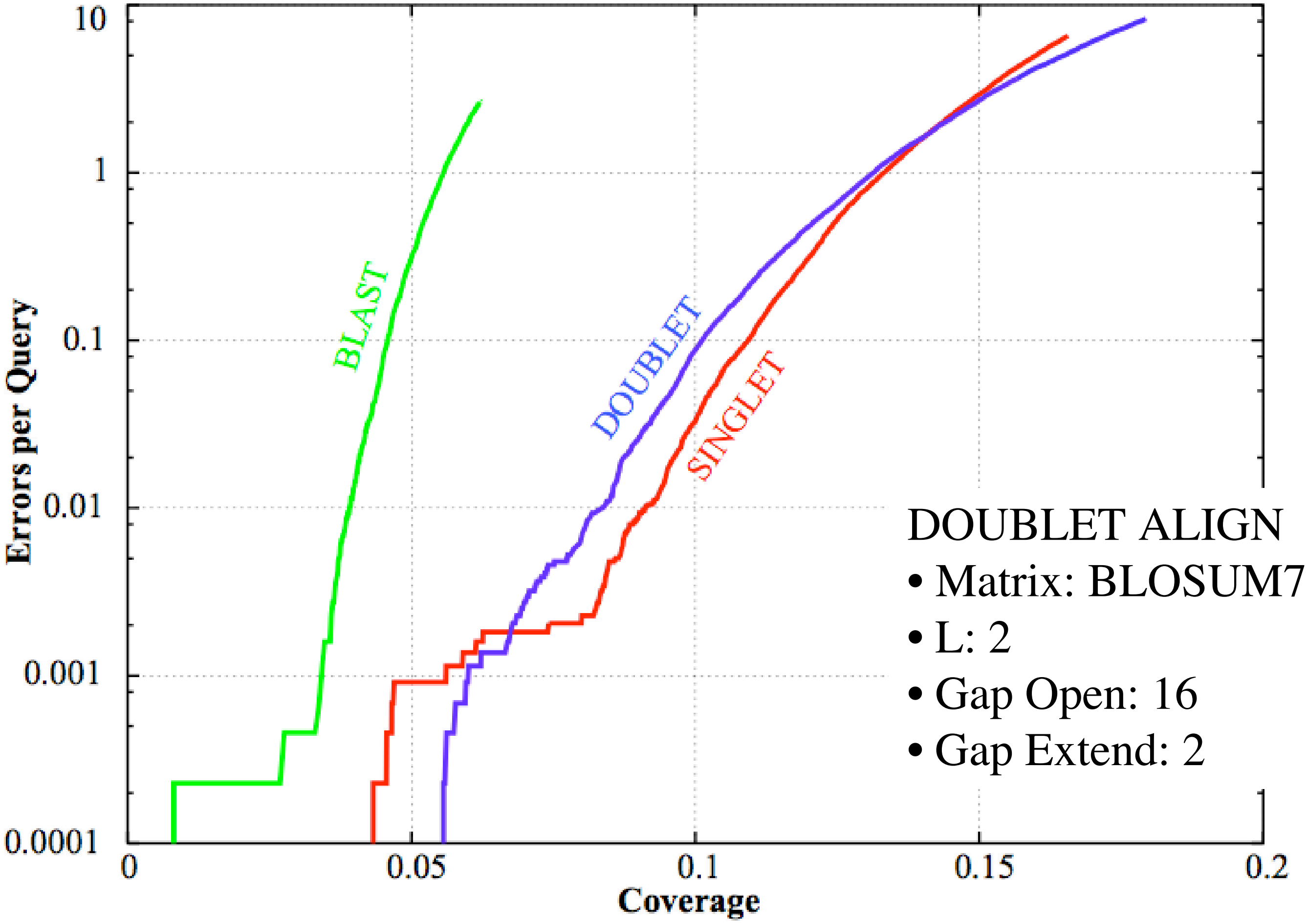
Parameter Space

- BLOSUM % Clustering
- Gap Penalty
  - Gap Open
  - Gap Extend
- Correlation Distance, L

Homology Detection: Coverage vs. Errors



Homology Detection: Preliminary Results



DOUBLET ALIGN

- Matrix: BLOSUM75-blocks13+
- L: 2
- Gap Open: 16
- Gap Extend: 2

Correlated Substitutions & Sequence Alignment

- Correlated Substitutions
  - Alignment score
  - Doublet matrixes
  - Efficient alignment algorithm
  - Evaluations: Remote homolog detection

- Effect of Correlations on Homology Detection:
  - Not much

- Future Work
  - Alignment quality?

Gavin E. Crooks <gec@compbio.berkeley.edu>