

Alignment

Procedure of comparing two (pairwise) or more (multiple) sequences by searching for a series of individual characters that are in the same order in the sequences

```
VLSPADKTNVKAAWGKVGAGHAGYEG
||| | | | | | |
VLSEGDWQLVLHVWAKVEADVAGEG
```

Sequence alignment

Bioinformatics Fall 2011

1

Sequence alignment

- Comparing DNA/protein sequences for
 - Similarity
 - Homology
- Prediction of function
- Construction of phylogeny
- Shotgun assembly
 - End-space-free alignment / overlap alignment
- Finding motifs

Sequence alignment

Bioinformatics Fall 2011

2

Kinds of alignment

- Local
 - Finding similar regions among
 - Dissimilar regions
 - Sequences of different lengths
- Global
 - Strings of similar size
 - Genes with a similar structure
 - Larger regions with a preserved order (syntenic regions)
- Multiple
 - Family of sequences provide a stronger signal

Sequence alignment

Bioinformatics Fall 2011

3

Multiple alignment example

```

-----VSLT PEEKAVTALMR VV VD --EYDIALRLLVY YF VTON
-----VSLT QEKAVLALMR VV VE --EYDIALRLLVY YF VTON
-----VLS PADKTVKAAMK YG AH AGEYDIALERFELS FF TTXT
-----VLS AADKTVKAAMK YG OH AGEYDIALERFELS FF TTXT
PVYDGGVAFLS AADKTVLSAMK FY OD YEDYDILLRFFETS TP AAEZ
-----VLS DGSQALVKSMEK VE AD YAGHAGDILLRFPK NP ETLZ
-----GALT ESQALVKSMEK FN AH IPRKTHRFFELVLEL AP AAKD

FFKSPDGLSTPDAMN PPKVAHGRKYLGFQDG --L AHDNL YG TFAT--LSELGCKLMD
FFKSPDGLSDPGLMKA PPKVAHGRKYLSPQEG --V HRLDL YG TFAA--LSELGCKLMD
YFHF--ELSE---GS AQWGRGRKVDALTKA --V AHVDEN PM ALSA--LSELGCKLMD
YFHF--ELSE---GS AQWGRGRKVDALTKA --V GRDGL PM ALSH--LSELGCKLMD
FFKFKGLTADLXKS ADVFMAERLIDVDA --V ASMDGT EF HSMEDLSGKAASFVD
KFRFKHLTEAMKAS EDLKKHVTVLTALJAI --L KXGSH EA ELAP--LAGSHTKNTIP
LSSFLAKGTSSEVPQR PELGAGAGVFKLYEAA TQE EYGVV AS DATLNLGSGVYSKVIYA

PENFLLGHWLVGVLAIR POREFTPPQA AYQKVVAGVANALA HKYH-----
PENFLLGHWLVVFLAIR PQRDFTPEGA SYQKVVAGVANALA HKYH-----
PWFKLLSRLSLFLAIR LPAEFTPAHA SLKFLSLSVTVLT SKYR-----
PWFKLLSRLSLFLAIR LPDFTPAHA SLKFLSLSVTVLT SKYR-----
PEYFKVLAIVADIVAAQ D-----A GFEKLLRMICILLR SAYT-----
IKYLEFTEALIVLKER RFGDPSADAG ANRQALELPRDTA ARYKELVDG
DAISFPYKALARTKRY VQWGRSEELG APTAVTEALVYK ---KQKDA-

```

Alignment between globins (human beta globin, horse beta globin, human alpha globin, horse alpha globin, cyanoaemoglobin, whale myoglobin, leghaemoglobin) produced by Clustal. Boxes mark the seven alpha helices composing each globin.

Sequence alignment

Bioinformatics Fall 2011

4

Homology

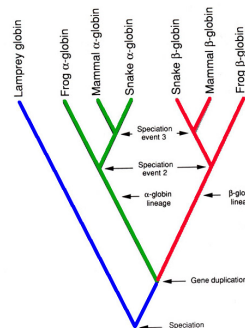
- Orthologs
 - Divergence follows speciation
 - Similarity can be used to construct phylogeny between species
- Paralogs
 - Divergence follows gene duplication
- Xenologs
 - Horizontal (inter-species) transfer of genes
- [Article on terminology](#)
- [inparalog and outparalog](#)
- [Preservation of function in homologs](#)

Sequence alignment

Bioinformatics Fall 2011

5

Orthologs and paralogs



Sequence alignment

Bioinformatics Fall 2011

6

Orthology / Paralogy

Lysosomal enzyme

● speciation
◆ duplication

Homology: two genes are homologous iff they have a common ancestor.

↔ *Orthology*: two genes are orthologous iff they diverged following a speciation event.

↔ *Paralogy*: two genes are paralogous iff they diverged following a duplication event.

⚠ Orthology ≠ functional equivalence

Sequence alignment Bioinformatics Fall 2011 7

Artifacts due to Paralogy

● speciation
◆ duplication

•Parallel phylogenies due to paralogy
•Gene loss can occur during evolution: even with complete genome sequences it may be difficult to detect paralogy

Sequence alignment Bioinformatics Fall 2011 8

Genetic variations

- Important for genetic variation and species fitness
 - Somatic versus gametic mutation
- Point mutations (change of a single base)
 - Chemical (environment effects)
 - replication error
 - transition (A-G or T-C, more common) or transversion
- Insertions or deletions (indels) can cause frame-shift
 - Transposable elements in DNA sequences (jumping genes)
 - Unequal crossing over
 - Replication slippage

Sequence alignment Bioinformatics Fall 2011 9

Genetic variations

- Duplication
 - a single gene (complete gene duplication)
 - part of a gene (internal or partial gene duplication)
 - Domain duplication
 - Exon shuffling
 - part of a chromosome (partial polysomy)
 - an entire chromosome (aneuploidy or polysomy)
 - the whole genome (polyploidy)

Sequence alignment Bioinformatics Fall 2011 10

Differing rates of evolution

- Functional/selective constraints (coding regions, promoter regions, other signals)
- Variation among different gene regions with different functions (different parts of a protein may evolve at different rates).
- Within proteins, variations are observed between
 - surface and interior amino acids in proteins (order of magnitude difference in rates in haemoglobins)
 - protein domains with different functions
 - regions which are strongly constrained to preserve particular functions and regions which are not

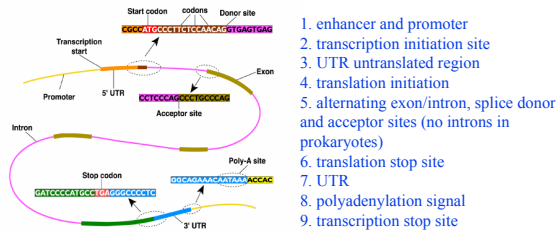
Sequence alignment Bioinformatics Fall 2011 11

Functional constraint

- Rates of substitution in different parts of the genome of mammals

Sequence alignment Bioinformatics Fall 2011 12

Structure of genes



Sequence alignment

Bioinformatics Fall 2011

13

Functional constraint

Average pairwise divergence among different regions of the human, mouse, rabbit, and cow beta-like globin genes

Region	Length of Region (bp) in Human	Average Pairwise Number of Changes	Standard Deviation	Substitution Rate (substitutions/ site/ 10 ⁹ year)
Noncoding, overall	913	67.9	14.1	3.33
Coding, overall	441	69.2	16.7	1.58
5' Flanking sequence	300	96.0	19.6	3.39
5' Untranslated sequence	50	9.0	3.0	1.86
Intron 1	131	41.8	8.1	3.48
3' Untranslated sequence	132	33.0	11.5	3.00
3' Flanking sequence	300	76.3	14.3	3.60

Note: No adjustment is made for the possibility that multiple changes may have occurred at some sites.

Sequence alignment

Bioinformatics Fall 2011

14

Synonymous vs. Nonsynonymous Substitutions

- Changes at the nucleotide level of coding sequence that do not change the amino acid sequence of a protein are called **synonymous substitutions**.
- In contrast, changes at the nucleotide level of coding sequence that do change the amino acid sequence of a protein are called **nonsynonymous substitutions**.
- Non-degenerate sites, k-fold degenerate sites

Sequence alignment

Bioinformatics Fall 2011

15

Synonymous vs. Nonsynonymous Substitutions

Divergence between different kinds of sites within the coding sequence of the human and rabbit beta-like globin genes.

Region	Number of sites (bp)	Number of Changes	Substitution Rate (substitutions/ site/ 10 ⁹ year)
Nondegenerate	302	17	0.56
Twofold degenerate	60	10	1.67
Fourfold degenerate	85	20	2.35

Sequence alignment

Bioinformatics Fall 2011

16

Common assumptions

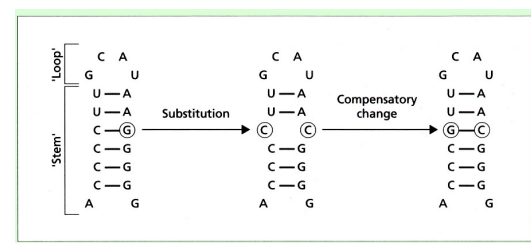
- All nucleotide sites change independently
- The substitution rate is constant over time and in different lineages
- The base composition is at equilibrium
- Most of these are not true in many cases...

Sequence alignment

Bioinformatics Fall 2011

17

Independence

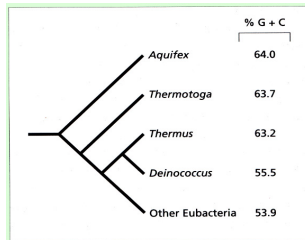


Sequence alignment

Bioinformatics Fall 2011

18

Base composition



Sequence alignment

Bioinformatics Fall 2011

19

Alleles

- Different forms of any given gene within a species of organism are known as alleles.
 - Determine phenotype
 - Dominant or recessive
- Changes in the relative frequencies of alleles during evolution.
- Genotype frequency
- Allele frequency

Sequence alignment

Bioinformatics Fall 2011

20

Example

genotype	A_1A_1	A_1A_2	A_2A_2
number	670	200	130
genotype frequency	$\frac{670}{1000}$	$\frac{200}{1000}$	$\frac{130}{1000}$
	0.67	0.20	0.13

$$\text{frequency of } A_1 = 0.67 + \frac{1}{2}(0.20) = 0.77$$

$$\text{frequency of } A_2 = 0.13 + \frac{1}{2}(0.20) = 0.23$$

Sequence alignment

Bioinformatics Fall 2011

21

- If only 6% of the population displays pale eyes (recessive gene e). What is the frequency of genotype Ee in this population?

$$q^2 = 0.06 \text{ ---> } q = 0.24$$

$$p + q = 1 \text{ ---> } p = 0.76$$

$$Ee = 2pq = 2(0.76)(0.24) = 0.36$$

Sequence alignment

Bioinformatics Fall 2011

22

Hardy-Weinberg equilibrium

- if p = frequency of allele A
q = frequency of allele a
- $p + q = 1$ or $p^2 + 2pq + q^2 = 1$
- Stability if only law of probability affects the frequency w/ which gametes combine to form new individuals

Sequence alignment

Bioinformatics Fall 2011

23

Assumptions

- Bisexual population
- Large population
- Random mating
- No mutation
- Migration ~ 0
- Natural selection does not occur

Sequence alignment

Bioinformatics Fall 2011

24

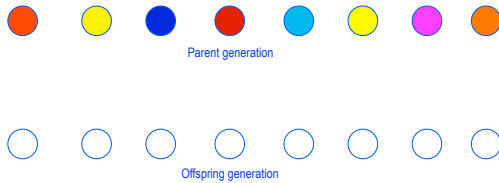
- A population that is in Hardy-Weinberg equilibrium will experience no change in either genotype frequency or allele frequency
- If one or more of the conditions is violated, genotype frequency and allele frequency will change

Neutral evolution

- Majority of mutations are selectively neutral.
- Most of the observed changes are a result of random genetic drift.
- In the absence of further mutation, eventually one allele will win out.

Genetic drift: Evolution without selection or mutation

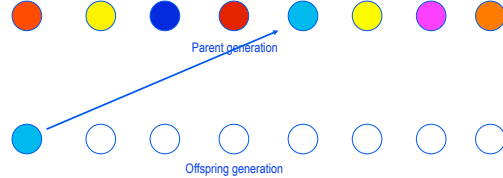
A population of fixed size. Each individual has its own type (color).
Individuals reproduce clonally.



Each individual has the same reproductive success on average:
Each offspring individual in the new generation has a parent chosen at random from the parent generation.

Genetic drift: Evolution without selection or mutation

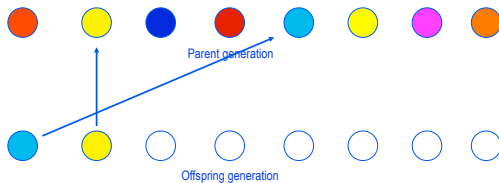
A population of fixed size. Each individual has its own type (color).
Individuals reproduce clonally.



Each individual has the same reproductive success on average:
Each offspring individual in the new generation has a parent chosen at random from the parent generation.

Genetic drift: Evolution without selection or mutation

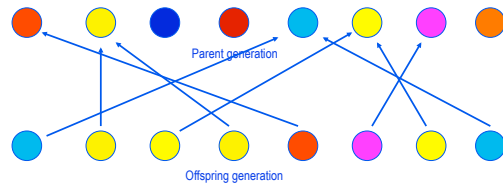
A population of fixed size. Each individual has its own type (color).
Individuals reproduce clonally.



Each individual has the same reproductive success on average:
Each offspring individual in the new generation has a parent chosen at random from the parent generation.

Genetic drift: Evolution without selection or mutation

A population of fixed size. Each individual has its own type (color).
Individuals reproduce clonally.



Each individual has the same reproductive success on average:
Each offspring individual in the new generation has a parent chosen at random from the parent generation.

Genetic drift: Evolution without selection or mutation

A population of fixed size. Each individual has its own type (color).
Individuals reproduce clonally.

Parent generation

Offspring generation

Each individual has the same reproductive success on average:
Each offspring individual in the new generation has a parent chosen at random from the parent generation.

Some have no offspring!

Sequence alignment Bioinformatics Fall 2011 **31**

Genetic drift: Eventually one color will take over

Allele fixation probability depends on initial ratio

2 N generations.

- For a clonally reproducing population of size N. After on average 2 N generations all but 1 lineage will go extinct.
- More complex for sexually reproducing entities but qualitatively the same idea: Almost all genetic material stems from a very small fraction of the ancestral population more than 2 N generations ago.

Sequence alignment Bioinformatics Fall 2011 **32**

Neutral evolution

- Loss of genetic variation within populations in the absence of mutations
 - Effect higher in small populations.
- Evolution is a result of random genetic drift and selection.
- Neutral evolution forms a null model for measuring effects of selection.
- Consider a species of size 1 million and a generation time of 2 years. Then fixation time is roughly
 - Neutral evolution alone: 2 million years
 - Selection and neutral evolution: 6,000 years with a selective advantage of 1%

Sequence alignment Bioinformatics Fall 2011 **33**

AS EXEMPLIFIED BY THE TYPE OF ALCOHOL DRINKERS

SO LONG AS IT ISN'T BAD

ONLY THE BEST

悪いものでなければ

良いものだけを

POISON

Smiff

＜中立言説＞ THE NEUTRAL THEORY

＜ダーウィンの進化論＞ DARWINIAN THEORY

Asah Shimbun

Sequ **34**

DNA versus Protein sequence comparison

	score	E(DNA)	E(prot)	E(its)
MUSGST	5090	10 ⁻⁵³	10 ⁻⁹⁹	10 ⁻¹²⁸
MUSGSTA	3655	10 ⁻⁶⁰	10 ⁻⁷³	10 ⁻¹²⁸
IRMSGSTAB	1930	10 ⁻³⁴	10 ⁻⁴⁸	10 ⁻¹¹⁸
MAAMGLUTRA	399	10 ⁻¹¹	10 ⁻⁷³	10 ⁻¹¹¹
RAGSYTYD	290	10 ⁻¹¹	10 ⁻²⁴	10 ⁻¹⁸
BSGSTM4	390	10 ⁻¹¹	10 ⁻¹⁸	10 ⁻¹⁸
RAUSTSY	372	10 ⁻¹⁰	10 ⁻⁷¹	10 ⁻¹⁸
BSGSTMB	358	10 ⁻⁹	10 ⁻⁴³	10 ⁻¹⁸
BSGSTMB3	322	10 ⁻⁷	10 ⁻²⁷	10 ⁻¹⁸
FEGST	240	0.00013	10 ⁻¹⁸	10 ⁻¹⁸
BSGSTPI	237	0.00049	10 ⁻¹⁷	10 ⁻¹⁸
MEGSTF	196	0.041	10 ⁻⁴	10 ⁻¹⁸
CRUGSTP	196	0.043	10 ⁻¹⁸	10 ⁻²¹
CRUGSTPE	196	0.044	10 ⁻¹⁸	10 ⁻²¹
IRMSGSTPE	191	0.13	10 ⁻¹⁸	10 ⁻²¹
BTRNAKOR	184	0.17	> 10	> 5
HUMGL2	170	0.29	> 10	> 5
IRMSGSTC1	170	0.67	10 ⁻⁵	> 5
IRMSGL11	168	1.0	10 ⁻⁷	> 5
MEGSTPEP	164	1.2	> 10	> 5.0
HUMTRDPPI1	161	1.7	> 10	> 5

Expectation values for searches against DNA (score, E(DNA)), protein (E(prot)), and translated DNA (E(its)) databases. A mouse glutathione transferase cDNA sequence (MUSGST) was used to search either the primate cDNA (IRMSGSTAB), and mammalian cDNA (IRMSGSTC1) databases of the GenBank DNA sequence database for the DNA sequence comparisons. Protein expectations (E(prot)) were calculated from a search the translated cDNA sequence against the GenPept sequence database, which includes all of translated GenBank. Untranslated sequences are italicized. E(prot) for italicized sequences are >= 100.

Comparing protein sequences is better

Sequence alignment Bioinformatics Fall 2011 **35**

Twilight zone of proteins

The Limits of Sequence Similarity

Observed Percent Identity

Evolutionary Distance (PAMs)

Twilight Zone

Pearson, '00 tutorial

Sequence alignment Bioinformatics Fall 2011 **36**

Differing rates of protein evolution

Table 4. Rates of change in protein families

Protein	Rate*	Protein	Rate
Fibrinopeptides	96	Thyrotropin beta chain	7.4
Growth hormone	37	Parathyria	7.3
Ig kappa chain C' region	37	Parathyroid	7.0
Kappa chain	33	BP11 Prostate inhibitors	6.2
Ig gamma chain C' region	31	Tropin	5.9
Luteinizing beta chain	30	Melanotropin beta	5.6
Ig lambda chain C' region	27	Alpha crystallin A chain	5.0
Complement C3a	27	Fallopian	4.8
Lactalbumin	27	Cytochrome b ₅	4.5
Fetal growth factor	26	Insulin	4.4
Somatotropin	25	Calcitonin	4.3
Pancreatic ribonuclease	21	Neurophysin 2	3.6
Liponase beta	21	Plasminogen	3.5
Hemoglobin alpha chain	20	Lactate dehydrogenase	3.4
Sebum albumin	19	Adenylate cyclase	3.2
Phospholipase A ₂	19	Triphosphatase esterase	2.8
Protease inhibitor PST1 type	18	Vasocinonin intestinal peptide	2.6
Protein	17	Carcinoma	2.5
Pancreatic hormone	17	Glyceraldehyde 3-P DH	2.2
Carbonic anhydrase C	16	Cytochrome C	2.2
Luteinizing alpha chain	16	Plat. ferritin	1.9
Hemoglobin alpha chain	12	Collagen	1.7
Hemoglobin beta chain	12	Tropomyosin C, skeletal muscle	1.5
Lipid-binding protein A-II	10	Alpha crystallin B-chain	1.5
Globin	9.8	Gibberin	1.2
Animal lysozyme	9.8	Gluconate DH	0.9
Myoglobin	8.9	Histone H2B	0.9
Insulin A ₁	8.7	Histone H2A	0.5
Nerve growth factor	8.5	Histone H3	0.14
Acid phosphatase	8.4	Ubiquitin	0.1
Myelin basic protein	7.4	Histone H4	0.1

ISMB tutorial,
Pearson 2000

*percent/100 My
From Nei, 1987; Doolittle et al., 1978

Sequence alignment

Bioinformatics Fall 2011

37

Molecular Clock Hypothesis

- A given gene or protein undergoes a constant rate of molecular substitution. (Zuckerkanndl and Linus Pauling)
 - The molecular clock may run at different rates in different proteins, but the number of differences between two homologous protein is correlated with the amount of time since speciation caused them to diverge independently.
- Does not hold for all proteins.
 - Different selective pressures
 - Different generation times

Sequence alignment

Bioinformatics Fall 2011

38

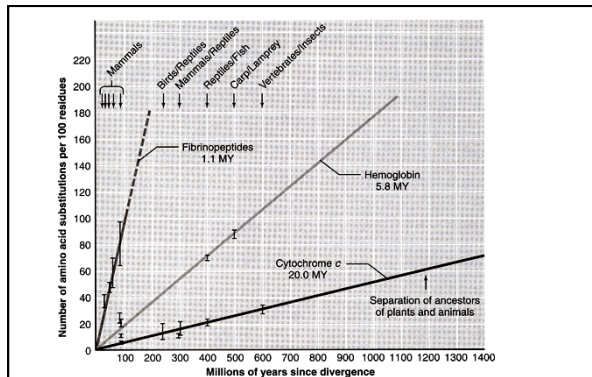


FIGURE 3.6 Numbers of amino acids replaced and species divergence times are well correlated for a number of proteins.

Sequence alignment

Bioinformatics Fall 2011

39

Estimating DNA Substitution Numbers

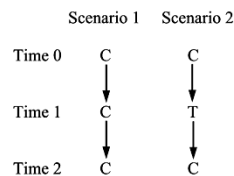
- The number of substitutions (K) observed in an alignment between two sequences is typically the single most important variable in any molecular evolution analysis.
- If an optimal alignment suggests that relatively few substitutions have occurred between two sequences, then a simple count of the substitutions is usually sufficient to determine a value for K .
- Alignments between sequences with many differences might cause a significant **underestimation** of the actual number of substitutions since the sequences last shared a common ancestor.

Sequence alignment

Bioinformatics Fall 2011

40

Jukes-Cantor Model



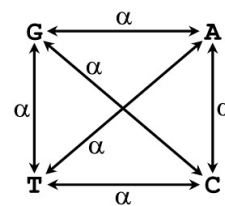
There are no guarantees that a particular site has not undergone multiple changes. Multiple substitutions at a single site would lead to underestimation of the number of substitutions that had occurred if a simple count were performed.

Sequence alignment

Bioinformatics Fall 2011

41

Jukes-Cantor Model



Apply Markov Chain analysis.

Q = transition matrix.

Q_{ij} = prob from state j to i .

(column normalized)

Initial state $s = (1/4, 1/4, 1/4, 1/4)$.

Q_s = state after time 1.

Q_s^t = state after time t .

All nucleotides changed to each of the three alternative nucleotides at the same rate, α .

Sequence alignment

Bioinformatics Fall 2011

42

What does Q^t look like?

using spectral expansion

$$\begin{bmatrix} \frac{1}{4}(1+3(1-4\alpha)^t) & \frac{1}{4}(1-(1-4\alpha)^t) & \frac{1}{4}(1-(1-4\alpha)^t) & \frac{1}{4}(1-(1-4\alpha)^t) \\ \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ \end{bmatrix}$$

Sequence alignment

Bioinformatics Fall 2011

43

Spectral Decomposition of Q

- Eigenvectors/eigenvalues of Q
 - $-u_1^T = [1/2, 1/2, 1/2, 1/2], \lambda_1 = 1$
 - $-u_2^T = [\sqrt{3}/2, -1/2\sqrt{3}, -1/2\sqrt{3}, -1/2\sqrt{3}], \lambda_2 = (1-4\alpha)$
 - $-u_3^T = [0, 2/\sqrt{6}, -1/\sqrt{6}, -1/\sqrt{6}], \lambda_3 = (1-4\alpha)$
 - $-u_4^T = [0, 0, 1/\sqrt{2}, -1/\sqrt{2}], \lambda_4 = (1-4\alpha)$
- Verify
 - Eigenvectors are orthonormal
 - $Au_i = \lambda_i u_i$

Sequence alignment

Bioinformatics Fall 2011

44

Spectral Decomposition of Q

- $Q = U \Lambda U^T = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \lambda_3 u_3 u_3^T + \lambda_4 u_4 u_4^T$
 $= \frac{1}{4} \mathbf{1} + (1-4\alpha) \begin{bmatrix} .75 & -.25 & -.25 & -.25 \\ -.25 & .75 & -.25 & -.25 \\ -.25 & -.25 & .75 & -.25 \\ -.25 & -.25 & -.25 & .75 \end{bmatrix}$

- $Q^n = U \Lambda^n U^T$
 $= \frac{1}{4} \mathbf{1} + (1-4\alpha)^n \begin{bmatrix} .75 & -.25 & -.25 & -.25 \\ -.25 & .75 & -.25 & -.25 \\ -.25 & -.25 & .75 & -.25 \\ -.25 & -.25 & -.25 & .75 \end{bmatrix}$

What happens as n becomes large?

Sequence alignment

Bioinformatics Fall 2011

45

Example

- Initial state = $[0.4, 0.2, 0.2, 0.2]$
- What happens after n steps?
- State = $[0.25, 0.25, 0.25, 0.25] + (1-4\alpha)^n [0.15, -0.05, -0.05, -0.05]$
- What is the effect of α ?

Sequence alignment

Bioinformatics Fall 2011

46

Estimating the actual rate of substitutions

- Let p be the observed rate of substitutions.
- Set $p = \frac{3}{4} (1 - (1-4\alpha)^t)$.
- $t = \ln(1-4p/3) / \ln(1-4\alpha)$
- $\ln(1-4\alpha) \approx -4\alpha$
- $3\alpha t = -\frac{3}{4} \ln(1-4p/3) =$ actual number of substitutions per site

Sequence alignment

Bioinformatics Fall 2011

47

An example

X: ACTTGTTGGATGATCAGCGGTCCATGCACCTGACAACGGT
 Y: ACATGTTGGTTGACCAGCGGTCCATGCGCCTGAGAACGGT

$p = 5/40 = 0.125$
 Number of substitutions per site = $-\frac{3}{4} \ln(1-4p/3) = 0.133$

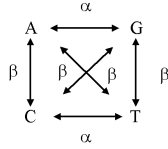
Sequence alignment

Bioinformatics Fall 2011

48

Other models of DNA evolution

- Kimura model
 - Transitions (changes among purines AG or among pyrimidines TC) are more common than transversions (changes between purines and pyrimidines).
- Position specific distributions



Sequence alignment

Bioinformatics Fall 2011

49

Score matrix

- Assign scores to each pair of symbol
 - Higher score means more similarity
- Compute odds (ratio of likelihood that two sequences are related as opposed to unrelated)
 - $P(x,y | M) / P(x,y | R)$
- Assume independence of sites
 - $\prod P(a,b | M) / P(a,b | R)$
- Take log (log-odds score)
 - $S(x,y) = \sum \log (P(a,b | M) / P(a,b | R)) = \sum s(a,b)$

Sequence alignment

Bioinformatics Fall 2011

50

Log-odds score

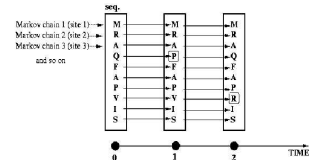
- $P(a,b | M)$ is the probability that symbols a,b derive from a common ancestor symbol, or occur as an aligned pair.
- $P(a,b | R)$ is the probability that symbols a,b are unrelated and occur as an aligned pair
 - Product of two frequencies
- Different ways of computing $P(a,b | M)$ for proteins
 - PAM
 - BLOSUM

Sequence alignment

Bioinformatics Fall 2011

51

PAM (Point Accepted Mutations)



- Developed by Dayhoff et al
- Based on a Markov model
- Independence of changes at different sites
- Higher score for accepted mutations
 - Implies a similar chemistry of residues
- A point can undergo more than one mutation.

Sequence alignment

Bioinformatics Fall 2011

52

PAM matrix construction

- Begin with 71 groups of related proteins (1572 changes)
 - At least 85% similarity within a group
- Obtain count of accepted mutations using parsimonious tree
- n_i = number of occurrences of amino acid i
 - n = total number of amino acids
- f_i = probability of occurrence of amino acid i
- $A_{ij} = A_{ji}$ = number of times i is aligned to j
- $A_j = \sum_i A_{ji}$
- $m_j = A_j / n_j$ = mutability of amino acid j

Sequence alignment

Bioinformatics Fall 2011

53

Construction of PAM1

- $M_{ji} = \lambda A_{ji} / n_j$
 - λ is a scaling constant
- $M_{jj} = 1 - \sum_{i \neq j} M_{ji} = 1 - \lambda m_j$
- Choose λ so that 1% of the amino acids undergo mutation.
 - $\sum_i \sum_{j \neq i} f_j M_{ij} = 0.01$, or $\lambda \sum_j f_j m_j = 0.01$
- $\text{score}(i,j) = 10 \log (M_{ij} / f_i f_j) = 10 \log (f_j M_{ij} / f_i f_j) = 10 \log ((f_j \lambda A_{ji} / n_j) / f_i f_j) = 10 \log ((\lambda A_{ji} / n) / f_i f_j)$
 - Note use of log odds
 - Why is the matrix score symmetric?
 - Is the matrix M symmetric?

Sequence alignment

Bioinformatics Fall 2011

54

PAM vs. BLOSUM

Equivalent **PAM** and **BLOSUM** matrices:

PAM100 = Blosum90
 PAM120 = Blosum80
PAM160 = Blosum60
 PAM200 = Blosum52
 PAM250 = Blosum45

BLOSUM62 is the usual default matrix to use.

Sequence alignment

Bioinformatics Fall 2011

61

Major Differences between PAM and BLOSUM

PAM	BLOSUM
Built from global alignments	Built from local alignments
Built from small amount of data	Built from vast amount of data
Counting is based on minimum replacement or maximum parsimony	Counting based on groups of related sequences counted as one
Better for finding global alignments and remote homologs	Better for finding local alignments
Higher PAM series means more divergence	Lower BLOSUM series means more divergence

Sequence alignment

Bioinformatics Fall 2011

62

Typical score matrix

- DNA
 - Match = +2
 - Mismatch = -3
 - Gap penalty = -5
 - Gap extension penalty = -2
- Protein sequences
 - Blossum62 matrix
 - Gap open penalty = -11
 - Gap extension = -1
- $\sum_{ij} p_i p_j s(i,j) = E(s(i,j)) < 0$, for $s(i,j)$ to be meaningful.

Sequence alignment

Bioinformatics Fall 2011

63

Measuring quality

- Ground truth
 - Positives (P)
 - Negatives (N)
- Ascertained truth
 - True Positives (TP)
 - True Negatives (TN)
 - False Positives (FP): *Type I error*
 - False Negatives (FN): *Type II error*
- $P = TP + FN$
- $N = FP + TN$

Sequence alignment

Bioinformatics Fall 2011

64

Measuring quality

- Sensitivity is the probability of a positive outcome when the answer should be positive = TP/P = Recall
- Specificity is the probability of a negative outcome when the answer should be negative = TN/N
- Precision = $TP/(TP+FP)$
- Receiver Operating Characteristics (ROC):
 - Plot of TP/P (sensitivity) vs. FP/N (1-specificity)
 - Random selection yields a line with slope 1.
 - (AUC) Area under ROC

Sequence alignment

Bioinformatics Fall 2011

65