Bioinformatics = Application of CS & IT to Bio & Med ∵ Large data size (⇒) ∵ Difficult computational problems (many disease & control seqs?) **CS:** String comparison (Identify genetic variants); **Stat:** How different are variant groups?; **Biomed:** Experimental validation & Functional study	**Adult:** 10^{14} cells, haploid genome (2 DNA copies), 3×10^9 nucleotides, 25000 protein-producing genes ← Data size **Why 2 copies?** ∵ 2^{23} combinations ∵ Error tolerance ∵ 1 can change in evolution
DNA: Nitrogenous base, Pentose sugar (ribose), Phosphate	Pyrimidine: CT / Purine: AG
Amino acid: Amine, Carboxylic acid, Side chain	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

Traditional approach: Hypothesis-driven, Bottom-up; Alternative approach: Data-driven, Top-down

Sequence Alignment and Searching

Assumption: Similar text strings have similar biological properties ∵ Diverge from common ancestor for short time ∵ Conservation suggests importance ∵ Similar structure → Similar functional units/domains

Given a sequence *r* (the mouse gene) and a database *D* of sequences (all human genes), find sequences *s* in *D* where sim(*r*, *s*) is above a threshold → Simplest way is to compute sim(*r*, *s*) for each s one by one

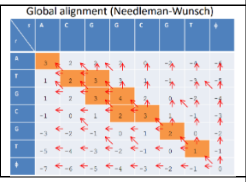
Definition: Given a set of sequences, an alignment is the same set of sequences with 0 or more gaps inserted into them so that → They all have same length → Each column has at least 1 non-gap

Good alignmt: Few mismatch/substitution and gap/indel → Optimal (highest score) ∴ Easy to compute similarities

2 seqs: Pairwise seq alignmt, >2 seqs: Multiple seq alignmt || Whole seq: Global alignmt, Parts of seq: Local alignmt

How? <10000 length: smart dynamic programming (≠ exponential alignmts), Long length: heuristic algorithms

Dynamic programming: Seqs *r*(*m*) & *s*(*n*): Make (m+1)×(n+1) optimal alignmt score table *V*(*i*,*j*) of suffix *r*[*i*..*m*] & *s*[*j*..*n*], $(r|s)[(m|n)+1]=\phi$ → Optimal score = *V*(1,1), $(m+1)(n+1)-1\approx mn$ alignmts, O(*mn*) polynomial ← Divide & Conquer: divide into smaller problems → solve small problems *r*[*i*..*m*] → combine results for original big problem ∵ Systematic: compare groups of alignmt simultaneously without needing to consider individual alignmts 1-by-1 ← Reuse sub-problems results: store & resue alignmt scores between suffixes


Global alignment (Needleman-Wunsch)

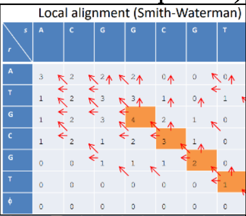
Scoring mtrx: DNA: Jukes-Cantor (eq prob), Kimura (transitn≠transversn); Protein: PAM (sub rate), BLOSUM (conserved seq blcks)

Gaps? Single nucleotide polymorphisms > indels ∴ gap penalty > mismatch, Small indels > large ∴ penalty ∝ gap size, Large gap > many small (same total) ∴ gap opening penalty > gap extension

Affine gap penalty: affine (straight line that may notpass origin), $y=-a-bx$; *y*: final gap score (-ve), *x*: gap size, *-a*: gap opening penalty, *-b*: gap size penalty

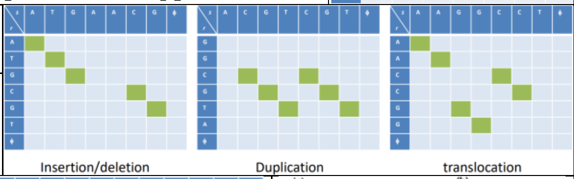
Without affine: gap penalty doesn't depend on other pos, With affine: depend on if it is last of gap (*-a*)

Local alignmt? Definition (⇑ +subseq) ∵ Similar inside domain ≠ outside → Output optimal subseq pairs


Local alignment (Smith-Waterman)

Heuristic (≠ optimal)**:** Find regions with high similarity by inspection & considering short subseqs → Combine & refine & results to get longer matches

Dot plot: insertion/deletion, duplication, translocation ☹ Must be exact match (mismatch need more computation) ☹ Large storage for plot ☹ Hard to determine resolution ☹ Not quantitative, mainly for visualisation

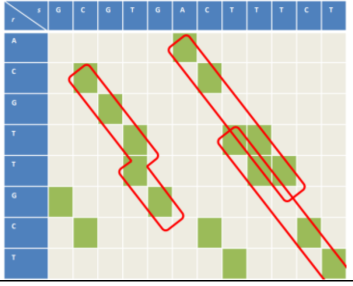

Insertion/deletion Duplication translocation

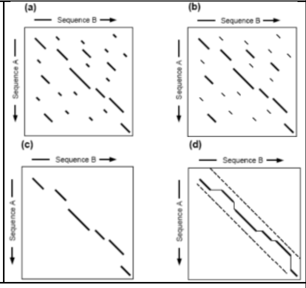
FASTA: Find *k* (protein: 1-2, DNA: 4-6) consecutive exact matches with simple scoring, build *k*-mer vs pos lookup table → Refine matches with formal substitution matrices → Combine matches allowing gaps, merge diagonals → Use banded DP on the matches

Miss optimal? ← Good non-exact local matches in step 1 ∵ large *k* ∵ High-scored mismatches (esp. protein) ← Many local candidates ∵ Only very best is chosen, discard rest

FASTA
```
>SEQ1
MTEITAAAA
>SEQ2
SATVSEIII
```

Space: (*n*-*k*+1) entries
Time: <O(*mn*)





BL vs FA: high-scoring inexact ∴ larger k, extend local matches rgdls presence of same diag match, evaluate stat sig of matched seqs

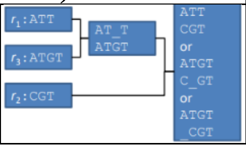
BLAST: Local exact & inexact similar matches → Extend adj char at ends until score < threshold → Stat sig E-value (exp. num in db) Nucl-nucl BL (blastn), Prot-prot BL (blastp), Nucl 6-frame translation-prot BL (blastx): 6FT on query → comp db prot seq, Prot-nucl 6FT BL (tblastn): comp query prot seq with 6FT nucl seq in db, Nucl 6FT-n6FT BL (tblastx): 6FT on query & db nucl seq → comp blastn if nucl conservation is expected (eg. ribosomal RNA), tblastx if prot conserevation is expected (eg. coding exons)

PSI-BLAST: Make similar seq profile (eg. CC[CG]C[AT][AT]T[GT]), BLAST again until no more new seq

Multiple Seq Alignmt: Make seq vs seq scoring matrix

```
seq1     2 ATG_AC 6
```

Clustal: Make dist matrix (dist = alignmt length – alignmt score) → Make tree → Align seqs using tree Clustal (ClustalW, ClustalX, Clustal Omega), T-Coffee, MAFFT, MUSCLE



Mutation Models and Molecular Phylogenetics

Evolutionary Distance: number of mutations between sequences/ time since divergence, $\mathrm{E}[\mathrm{K_{sup}}]$

Mutation Model: Prob mdl of mttn freq, What kind of mttn more feq? Assumption (usually not true but simpler): Sites are independent, Mttn rates are same for diff sites at diff time, Future states don't depend on past states

DNA: simple param
Prot: biochem prop
Prot sub < DNA sub

Jukes-Cantor: Eq rate of sub to other bases in 1 time, $\mathrm{P_{sub}}=\alpha$; $\mathrm{P_{same}}=1-3\alpha$; $\mathrm{P_{X\to X}}(t)=¼+¾e^{-4\alpha t}$; $\mathrm{P_{X\to Y}}(t)=¼-¼e^{-4\alpha t}$ $\mathrm{P_{same}}(t)=\mathrm{P_{X\to X}}(t)^2+3\mathrm{P_{X\to Y}}(t)^2=¼+¾\mathrm{e}^{-8\alpha t}=1-\mathrm{p_{diff}}$; $\alpha t=-⅛\ln(1-{}^4/_3\,\mathrm{p_{diff}})$

Estimate $\mathrm{p_{diff}}=\frac{x}{n}$, *x*: num of sites diff btwn obs seqs, *n*: seq length → $\mathrm{E}[\mathrm{K_{sup}}]=6\alpha t$; $\mathrm{Var}=\frac{\mathrm{p_{diff}}-\mathrm{p_{diff}}^2}{n(1-{}^4/_3\mathrm{p_{diff}})^2}=\frac{{}^x/_n-({}^\mathrm{x}/_\mathrm{n})^2}{n(1-{}^{4x}/_{3n})^2}$

Kimura 2-param: $\mathrm{P_{tsn}}=\alpha>\mathrm{P_{tvn}}=\beta$; $\mathrm{P_{same}}=1-\alpha-2\beta$; $\gamma=¼e^{-4\beta t}$; $\delta=½e^{-2(\alpha+\beta)t}$ $\mathrm{P_{X\to X}}(t)=¼+\gamma+\delta$; $\mathrm{P_{Xtsn}}(t)=¼+\gamma-\delta$; $\mathrm{P_{Xtvn}}(t)=¼-\gamma$

Transitn: pu↔pu A↔G py↔py C↔T
Transversn: py↔pu A↔C↔G↔T↔A

Estimate $\mathrm{p_{d1}}=\frac{x_1}{n}$; $\mathrm{p_{d2}}=\frac{x_2}{n}$, x_1: num of tsns, x_2: num of tvns → $\mathrm{E}[\mathrm{K_{sup}}]=½\ln(1-2\mathrm{p_{d1}}-\mathrm{p_{d2}})^{-1}+¼\ln(1-2\mathrm{p_{d2}})^{-1}$

$\mathrm{Var}=\frac{1}{n}\left(\mathrm{p_{d1}}\left(\frac{1}{1-2\mathrm{p_{d1}}-\mathrm{p_{d2}}}\right)^2+\mathrm{p_{d2}}\left(\frac{1}{2-4\mathrm{p_{d1}}-2\mathrm{p_{d2}}}+\frac{1}{2-4\mathrm{p_{d2}}}\right)^2-\left(\frac{\mathrm{p_{d1}}}{1-2\mathrm{p_{d1}}-\mathrm{p_{d2}}}+\frac{\mathrm{p_{d2}}}{2-4\mathrm{p_{d1}}-2\mathrm{p_{d2}}}+\frac{\mathrm{p_{d2}}}{2-4\mathrm{p_{d2}}}\right)^2\right)$; More acc for more divg seqs

PAM (Pt Accepted Mttn): Acptd=Survd; PAM*x* (prob of sub *i*→*j* given *x* sub per 100 aa) = PAM1^x; grps of related prots; asymetric

BLOSUM (BLOck of aa SUb Mtrx): local alignmt of conserved prot regions; BLOSUM*y* (local alignmt with seqs >*y*% identical)

Log-odd score: $S_{ij} = \frac{1}{\lambda}\log_2\frac{p_{ij}}{p_i p_j}$, p_{ij}: fraction of subs btwn aa *i* & *j*, $p_{i|j}$: fraction of sites with aa *i* | *j*, λ: scaling factor; symmetric

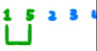
Newick: `((A:0.1,B:0.2)n?:0.3,C:0.4);` NEXUS: Map species to nums + Newick PhyloXML: XML-based

Phylogenetic Tree Reconstruction: Given *k* DNA/Prot seqs → Order of divg events (topology), Ancestral seqs (node seqs), Branch length (time since divg); Exponential, rooted: (2*k*-3)! topologies, unrooted: (2*k*-5)! tops; Sequences-based, exact seq: Parsimony, Maximum likelihood Distance-based, heuristic: UPGMA, Nghbr joining

UPGMA (Unwghtd Pair Grp Mthd with Arithmetic mean): Calc lowest avg pairwise dist → Merge clusters ⮌ Branch length =? Divg event count (tree layer count)

	s_{15}	s_2	s_3	s_4
s_{15}	0	8	6	6
s_2	8	0	2	6
s_3	6	2	0	10
s_4	6	6	10	0

$d(s_{15},s_2) = 5$
$d(s_{15},s_3) = 5.5$
$d(s_{15},s_4) = 4.5$
$d(s_2,s_3) = 2.5$
$d(s_2,s_4) = 5.5$
$d(s_3,s_4) = 6$

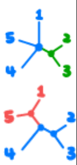


Neighbour Joining: Calc lowest $Q(i,j) = (r-2)\,d(C_i,C_j) - u(C_i) - u(C_j)$, r: current num of clusters, u: column sum, Branch length $= \frac{d_{ij}}{2} + \frac{|u_i-u_j|}{2(r-2)}$, Last node: remove hub & write dist

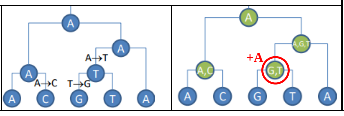
d	s_1	s_{23}	s_4	s_5
s_1	0	6	8	2
s_{23}	6	0	6	6
s_4	8	6	0	6
s_5	2	6	6	0

u	s_1	s_{23}	s_4	s_5
	16	18	20	14

Q	s_1	s_{23}	s_4	s_5
s_1	0	-22	-20	-26
s_{23}	-22	0	-26	-20
s_4	-20	-26	0	-22
s_5	-26	-20	-22	0



Maximum Parsimony: Assume: Tree with fewest mttns is correct, Independnt sites Large Prsmy: Given seqs → Rooted tree top (min mttn branch), Small Prsmy: Given seqs & tree → Ancestral seq (min mttn branch); Upward propagation → Downward



Maximum Likelihood: Maximise prob of obs data by a prob mdl givn mdl params $\Pr(X|\theta)$, X: obs data (alignd seqs), θ: mdl params Big likelihood (hard): θ: tree top, mttn rate, divg time; Small likelihood (gradient ascent): Given tree top, θ: mttn rate, divg time

Motifs and Domains

Motif/Domain: Patterns that Appear freqly (unlikely random/ over-represented), Known functional roles, Evolutionarily conserved Transcription Factor Binding Sites: 6-10bp DNA regulatory seqs that freqly appear at spec genomic locations, evolutny conserved Prot domains: similar subseq on diff prots with particular func, evolutny conserved; Domains > motifs, func/structural independence

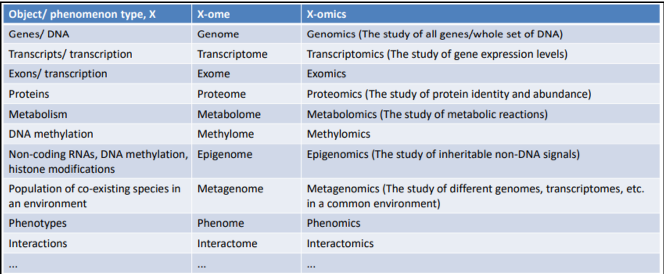
Representation? Exact rep: Consensus seq, Degenerate seq, Regex; Stat rep: Position weight matrix (probability base vs pos, +1 pseudo-count to all), Seq logo

R	Y	S	W	K	M	B	D	H	V	N	-
AG	CT	GC	AT	GT	AC	!A	!C	!G	!T	?	/

Pfam: Alignmt of rep seed seq, Profile HMM (prob mdl ≈PWM +pos relatshp, made from seed with HMMER3, used to scan prot seqs in UniProtKB), Alignmt of seqs above threshold score, Domain architecture, Phylogeneitc tree of seqs, Strcutral info Entries: Family (cllctn of reltd prots), Dmn (strc unit found in multpl contxts), Repeat (only stbl when multpl cps are present), Motif (short unit outside globular dmns); Clans: Seq, Structr, Profile HMM; Cmponts: Pfam A (high q, manually curated), Pfam B (l q, auto)

High-throughput Data Processing and Analysis

X-ome: Large amnt of data of X; **X-omic:** To study the data; **X-omics:** The area of studying the data; Omic Research: ☺ High-throughput, parallelisable, fast, less tedious, inexpensive ☺ Comprehensive ☺ Unbiased ☺ Easy to study interactions & combinatorial effects || ☹ Noise ☹ Secondary effects ☹ Lack of clear hypotheses ☹ High initial cost (machine)



Object/ phenomenon type, X	X-ome	X-omics
Genes/ DNA	Genome	Genomics (The study of all genes/whole set of DNA)
Transcripts/ transcription	Transcriptome	Transcriptomics (The study of gene expression levels)
Exons/ transcription	Exome	Exomics
Proteins	Proteome	Proteomics (The study of protein identity and abundance)
Metabolism	Metabolome	Metabolomics (The study of metabolic reactions)
DNA methylation	Methylome	Methylomics
Non-coding RNAs, DNA methylation, histone modifications	Epigenome	Epigenomics (The study of inheritable non-DNA signals)
Population of co-existing species in an environment	Metagenome	Metagenomics (The study of different genomes, transcriptomes, etc. in a common environment)
Phenotypes	Phenome	Phenomics
Interactions	Interactome	Interactomics
…	…	…

Omic workflow: Data production → Dt processing (QC, dt normalisation) → Dt analysis (pattern discovery) → Dt annotation & comp (evaltn of stat sig) → Selctn & sumrstn of results → Hypothesis formation → Exprmntl vldtn

Sanger sequencing: low-throughput, high reliability, 1000 nucl per reaction Parallel sequencing: platform (mobile, solid phase), immobilisation (primer, template, polymerase), longer reads, high error rate, high cost, single-cell sequencing

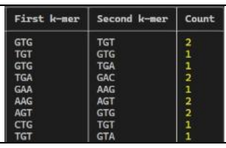
Shotgun sequencing: Cut long DNA randomly into short frags with high coverage overlap → Rate quality score of each read → Seq assembly (*de novo* asmbly)/ alignmt (re-seqcing)

FASTQ:

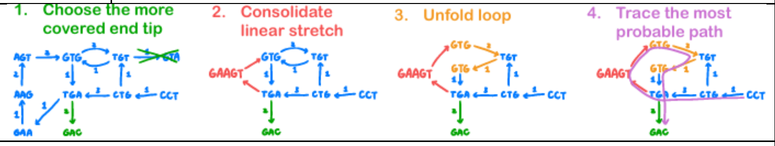
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTAT
+
!''*((((***+))%%%++)(%
```

↑ Quality (Phred score)
= ASCII code – 33 = $-\log_{10}$ P(error)

de Bruijn graph: Make *k*-mer subseq adj table (1st + 2nd subseq vs count) → graph → reconstruction Error? ← Tips, Bubbles, Low-coverage paths



First k-mer	Second k-mer	Count
GTG	TGT	2
TGT	GTG	1
GTG	TGA	1
TGA	GAC	2
GAA	AAG	1
AAG	AGT	2
AGT	GTG	2
CTG	TGT	1
TGT	GTA	1


1. Choose the more covered end tip
2. Consolidate linear stretch
3. Unfold loop
4. Trace the most probable path
AGT
GTG
TGT
GTA
AAG
GAA
TGA
CTG
CCT
GAC
GAAGT

CIGAR: **M**atch, **S**ubstitution, **I**nsertion, **D**eletion

Measure Gene Expression Level, high-throughput: Microarrays (design probes, hybridisation, fluorescent dye), cDNA (RNA-seq) Microarray: noisy (cross-hybridisation, background signal, sensitive to exp condition), don't know source gene if not unique RNA-seq: better S/N ratio, wide signal range, no need prior seq knowledge, don't know source gene if not unique

Two-way Hierarchical Clustering: Eucledian dist (if abs exp lvl matter), Pearson correlation (if only trend matter) $r(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$

***K*-means:** Partition into *k* clusters, unsupervised Random representative ⟼ Cluster nearest → New cluster rep (centroid of cluster) → Decluster, ↺ until stabilise

Functional Annotations

Func Genomic Elmnts/ Biotypes: Prot-coding genes (transcript, exon, intron, coding seq, untranslated region), Non-coding RNA

GFF/GTF: seqname, source (gen prgm?), feature (codon?), start, end, score, strand, frame, group

Ontology: The philphcal stdy of the nature of being, existence, or reality as such, as well as the basic catgrs of being and their relatns

GO: Sub-O: MF (lo-lvl func), BP (hi-lvl proc), CC (where? found); Part: Dirctd acyclic grph (is-a, part-of, reglats), Orgsm-spec instnc

KEGG: Metabolic pthwy, Genetic info procsng, Envrnmtal info procsng, Cellular proc, Orgsm sys, Human disease, Drug dev

Functional Enrichment: Test for co-expression with null hypothesis, Correlation ≠ Causation ≠ Related

Molecular Structures

Primary (seq), Secondary (local), Tertiary (global), Quaternary (multiple molecular interaction)

CATH hierarchy: Class (comp of sec struct), Architecture (shape), Topology (connection of sec structs), Homologus (with common ancestor)