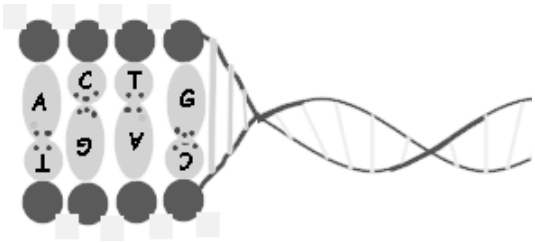
Bio2

Gene and Protein Prediction

Gene prediction

- What is a gene?
  - Simple definition: A stretch of DNA that encodes a protein and includes the regulatory sequences required for temporal and spatial control of gene transcription.
- Characteristics of genes.
  - What genetic features can we use to recognise a gene?

DNA structure



Bases: A,C,G and T

Chemically, A can only pair with T and G with C

Two strands, 5' and 3' Genes are encoded along one side of the DNA molecule. The 5' end being at the left hand side of the gene.

Codons and ORFs

- Three bases that encode an amino acid or stop site.
- A run of valid codons is an Open Reading Frame.
- An ORF usually starts with a Met
- Ends with a nonsense or stop codon.

Second base of codon

First base of codon	U	C	A	G	Third base of codon
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU SER UCC SER UCA SER UCG SER	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thy ACC Thy ACA Thy ACG Thy	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu CAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Predicting ORFs

- 64 total codons
- 3 stop codons, 61 codons for amino acids
- Random sequence 1:21 ratio for stop:coding.
- = 1 stop codon every 63 base pairs
- Gene lengths average around 1000 base pairs.

Finding ORFs

- One algorithm slides along the sequence looking stop codons.
- Scans back until it finds a start codon.
- Fails to find very short genes since it it looking for long ones
- Also fails to find overlaping ORFs
- There are many more ORFs than genes

Amino Acid Bias

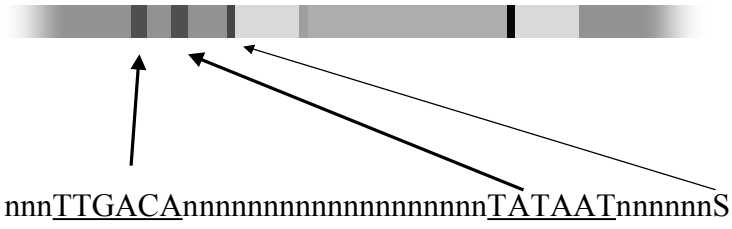
- The amino acids in proteins are not random
  - leucine has 6 codons
  - alanine has 4 codons
  - tryptophan has 1 codon
- The random the ratio would be 6:4:1
- In proteins it is 6.9:6.5:1
  - i.e. it is not random

Gene Prediction

- Take all factors into consideration
- Prokaryotes
  - No Nucleus
  - 70% of the genome encodes protein
  - No introns

Prokaryote gene structure

1. Promoter region



nnnTTGACAnnnnnnnnnnnnnnnnnnnTATAATnnnnnnS

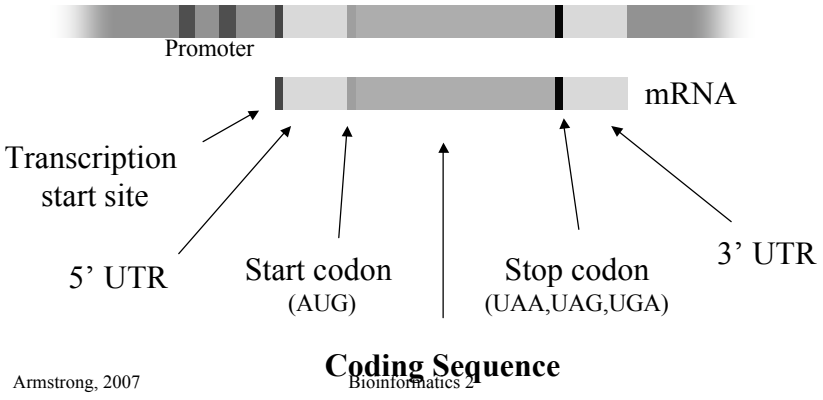
(consensus sequence for *E.coli*.)

Probability matrix for TATA box

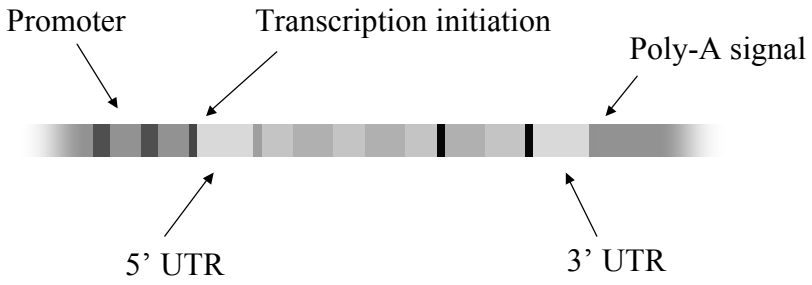
Pos:	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

Prokaryote gene structure

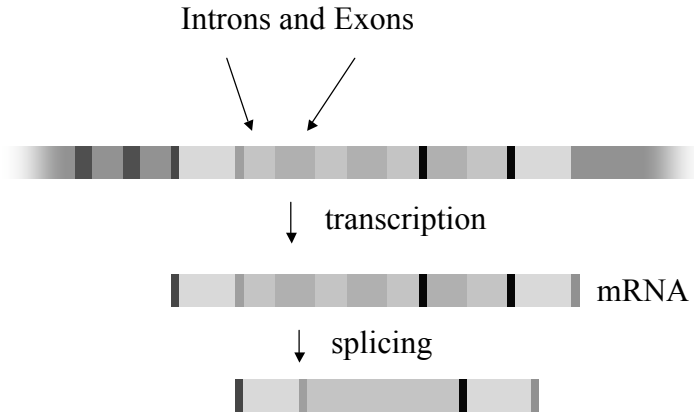
2. Transcribed region (mRNA)



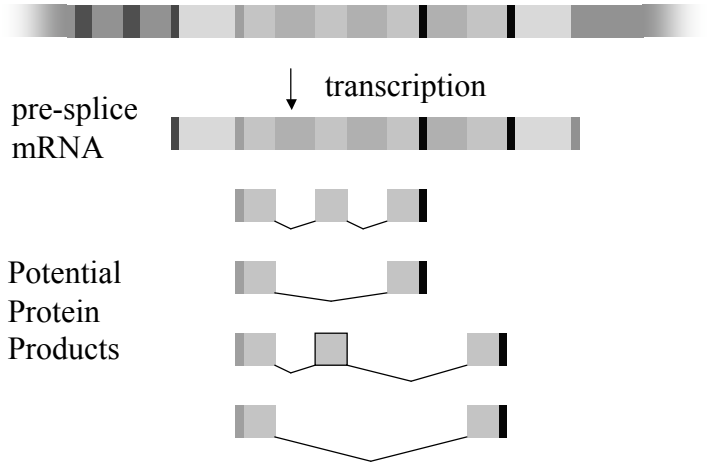
Eukaryote gene structure



Eukaryote gene structure

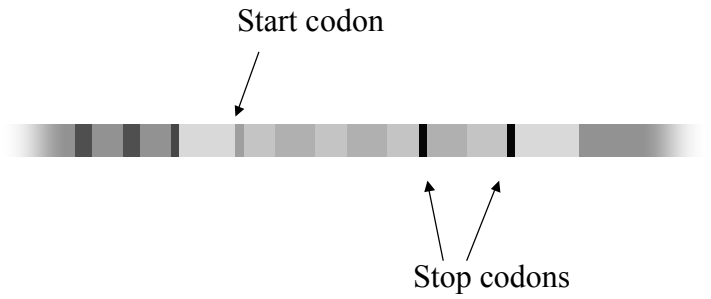


Functional significance of Introns and Exons



1 gene - 4 protein products

Eukaryote gene structure



Intron/Exon structure allows multiple start and stop codons

HMMs for codons

- Model based on examining 6 consecutive bases (i.e. all three reading frames).
- Based on statistical differences between coding and non coding regions
- 5^{th} order Markov Model.
- Given 5 preceding bases, what is the probability of the 6^{th}?
- Homogenous model (ignores reading frame)

HMMs for codons

- Homogenous models have two tables, one for coding, one non coding.
- Each table is has 4096 entries for the potential 6 base pair sequences
- Non-homogenous models have three tables for possible reading frames
- Short exons cause these models problems
- Hard to detect splice sites

Glimmer

- Uses non-homogenous HMMs to predict prokaryote gene sequences
- Identifies ORFs
- Trains itself on a prokaryote genome using ORFs over 500 bp
- http://www.cs.jhu.edu/labs/compbio/glimmer.html

Predicting Splice Sites

- There are some DNA features that allow splice sites to be predicted
- These are often species specific
- They are not very accurate.

NetGene2

- Neural network based splice site prediction
- Trained on known genes
- Claims to be 95% accurate
- Human, C. elegans & Arabidopsis thaliana
- http://www.cbs.dtu.dk/services/NetGene2/

HMMgene

- Based on an HMM model of gene structure
- Predicts intron/exon boundries
- Predicts start and stop codons
- Known information can be added (e.g. from ESTS etc)
- Outputs in GFF format

GFF Format

- Exchange format for gene finding packages
- Fields are:
  - <seqname> name, genbank accession number
  - <source> program used
  - <feature> various inc splice sites
  - <start> start of feature

GFF Format

  - <end> end of feature
  - <score> floating point value
  - <strand> +, - (or .. for n/a)
  - <frame> 0,1 or 2

GenScan

- Probabilistic model for gene structure based on a general HMM
- Can model intron/exon boundries, UTRs, Promoters, polyA tails etc
- http://genes.mit.edu/GENSCAN.html

Given a new protein sequence…

- What is the function?
- Where is the protein localised?
- What is the structure?
- What might it interact with?

Given a new protein sequence…

- What is the function?
- Have we seen this protein or a very similar one before?
  - If yes then we can infer function, structure, localisation and interactions from homologous sequence.
- Are there features of this protein similar to others?

Protein Families

- Proteins are complex structures built from functional and structural sub-units
  - When studying protein families it is evident that some regions are more heavily conserved than others.
  - These regions are generally important for the structure or function of the protein
  - Multiple alignment can be used to find these regions
  - These regions can form a signature to be used in identifying the protein family or functional domain.

Protein Domains

- Evolution conserves sequence patterns due to functional and structural constraints.
- Different methods have been applied to the analysis of these regions.
- Domains also known by a range of other names:

motifs patterns prints blocks

Profiles

- Given a sequence, we often want to assign the sequence to a family of known sequences
- We often also want to assign a subsequence to a family of subsequences.

Profiles

- Examples include assigning a gene/protein to a known gene/protein family, e.g.
  - G coupled receptors
  - actins
  - globins

Profiles

- Also we may wish to find known protein domains or motifs that give us clues about structure and function
  - Phosphorylation sites (regulated site)
  - Leucine zipper (dna binding)
  - EGF hand (calcium binding)

Creating Profiles

- Aligning a sequence to a single member of the family is not optimal
- Create profiles of the family members and test how similar the sequence is to the profile.
- A profile of a multiply aligned protein family gives us letter frequencies per column.

Matching sequences to profiles

- We can define a distance/similarity cost for a base in each sequence being present at any location based on the probabilities in the profile.
- We define define costs for opening and extending gaps in the sequence or profile.
- Therefore we can essentially treat the alignment of a sequence to a profile as a pairwise alignment and use dynamic programming algorithms to find and score the optimal alignments.

Protein profiles

- Multiple alignments can be used to give a consensus sequence.
- The columns of characters above each entry in the consensus sequence can be used to derive a table of probabilities for any amino acid or base at that position.

Protein profiles

- The table of percentages forms a profile of the protein or protein subsequence.
- With a gap scoring approach - sequence similarity to a profile can be calculated.
- The alignment and similarity of a sequence / profile pair can be calculated using a dynamic programming algorithm.

Protein profiles

- Alternative approaches use statistical techniques to assess the probability that the sequence belongs to a family of related sequences.
- This is calculated by multiplying the probabilities for amino acid *x* occurring at position *y* along the sequence/profile.

Probabilistic models

- Protein sequences are over 300 ave length.
- Random amino acid probability is 0.05
- Multiplying low probabilities together can cause underflow errors.
- Move into log space:
  - Take the log of the probabilities and sum.

HMMs

- A hidden markov model (HMM) is a refinement of this approach:
- HMMs can be visualised as finite state machines with a begin and an end state.
- FSMs move through a series of state emitting some kind of output report either at the end or during a transition from one state to another.

Protein profile HMMs

- In the profile of a protein sequence, there are effectively 3 states the model can be in:
  - 1. Match (exact or substitution)
  - 2. Insertion
  - 3. Deletion

Scoring profile HMMs

- The score of a sequence is the product of the probabilities that describe the path taken through the model used to recreate the sequence.
- Again, a log transformation allows the log of the probabilities to be summed rather than the probabilities multiplied.

Tools for HMM profile searches

- Meme and Mast at UCSD (SDSC)
- http://meme.sdsc.edu/
- MEME
  - input: a group of sequences
  - output: profiles found in those sequences
- MAST
  - input: a profile and sequence database
  - output: locations of the profile in the database

Summary

- Multiple alignment is used to define and find conserved features within DNA and protein sequences
- Profiles of multiply aligned sequences are a better description and can be searched using pairwise sequence alignment.
- Many different programs and databases available.

Secondary Databases

- PDB
- Pfam
- PRINTS
- PROSITE
- ProDom
- SMART
- TIGRFAMs

PDB

RCSB PDB PROTEIN DATA BANK

- Molecular Structure Database
- Contains the 3D structure coordinates of 'solved' protein sequences
  - X-ray crystallography
  - NMR spectra
- 29429 protein structures

Superfamily 1.65

HMM library and genome assignments server

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known (based on PDB)

Pfam

- Database of protein domains
- Multiple sequence alignments and profile HMMs
- Entries also annotated
- Swiss-Prot DB all pre-searched
- New sequences can be searched as well.
  - 7973 entries in Pfam last update

- Database of ‘protein fingerprints’
- Group of motifs that combined can be used to characterise a protein family
- ~11,000 motifs in PRINTS DB
- Provide more info than motifs alone

‘linear’ motifs

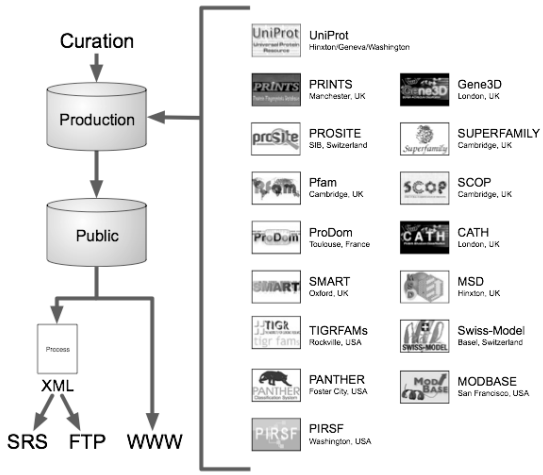
- Not all protein motifs are easy to find
- Linear motifs involved in protein-protein interactions
  - Very degenerate
  - Found in specific regions of proteins
  - Require special treatment
  - Neduva *et al*, PLOS 2005

Linking it all together…

- Database Searches
  - Multiple Alignments
  - Find known motifs and domains
  - Find possible similar folds
- Prediction algorithms
  - Properties of amino acids
  - Predicting folding
  - Finding cysteine bonds

InterPro

- EBI managed DB
- Incorporates most protein structure DBs
- Unified query interface and a single results output.



See http://www.ebi.ac.uk/interpro/

InterPro

DATABASE	VERSION	ENTRIES
SWISS-PROT	48	197228
PRINTS	38	1900
TREMBL	31.1	2342938
PFAM	18	7973
PROSITE	19.10	1882

Currently 15 databases, plans to add 3 new ones this month.

PredictProtein

http://www.embl-heidelberg.de/predictprotein/

Database searches:

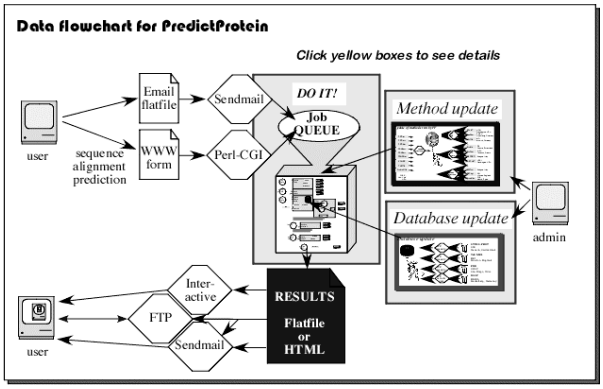
- generation of multiple sequence alignments (MaxHom)
- detection of functional motifs (PROSITE)
- detection of composition-bias (SEG)
- detection of protein domains (PRODOM)
- fold recognition by prediction-based threading (TOPITS)

PredictProtein

Predictions of:

- secondary structure (PHDsec, and PROFsec)
- residue solvent accessibility (PHDacc, and PROFacc)
- transmembrane helix location and topology (PHDhtm, PHDtopology)
- protein globularity (GLOBE)
- coiled-coil regions (COILS)
- cysteine bonds (CYSPRED)
- structural switching regions (ASP)

Data and methods in PredictProtein



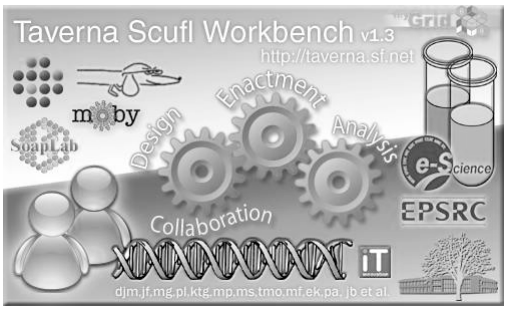
Add data and programs run at central site and updated on a regular basis

Too many programs/databases

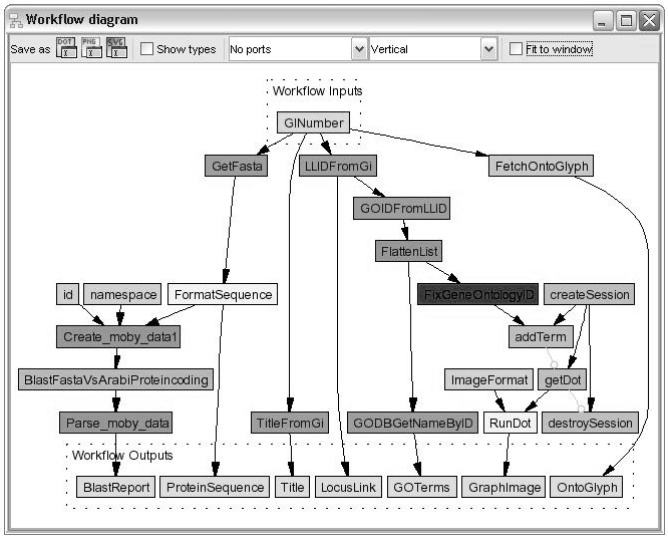
- How do we keep track of our own queries?
  - Repeat an old query
  - Run the same tests on a new sequence
  - Run 100s of sequences..
  - Document the process for a paper or client or for quality assurance

Workflow managers

- Locate and manage connections to software and databases
- Record actions
- Replay a workflow at a later date or against multiple sequences
- Manages redundant external sources (e.g. multiple blast servers)
- Can connect to specialist local sources



- http://taverna.sourceforge.net/
- Open source and free to download
- Runs on PC/linux/mac
- Drag-n-Drop interface to bioinformatics analysis



Example workflow from on-line taverna documentation

Given a new protein sequence…

- *What is the function?*
- *Where is the protein localised?*
- *What is the structure?*
- *What might it interact with?*

These are not fully solved problems. The latest issue of Bioinformatics (today) contains many new studies and tools addressing these problems.

Protein-Ligand interactions

- Most proteins do not act alone
- Most interact with other molecules
  - Other proteins
  - Small molecules
  - Drugs
- The shape and amino acid composition come together to form the site of interaction.
- 'Grand Challenge' in Bioinformatics: Can we accurately predict if two molecules will interact with each other based on sequence alone?