Methods for examining high-throughput biological data

Yee Hwa (Jean) Yang

Sept 26 2010

Open access, freely available online

Essay

Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better

Joel E. Cohen

Published in 2004 PLoS Biology

- Current biology generate a large amount of data from high-throughput technologies. Drawing biologically meaningful inferences from these data requires analysis in the framework of good mathematical models.
  - Mathematics has become a necessary tool for biology.

- Currently available computer power allows us to investigate sufficiently detailed mathematical models to draw biologically realistic inferences.
  - Mathematics has become a useful tool for biology.

The Past

Subject	Illustrative Reference
Age structure of stable populations	Euler 1760
Logistic equation for limited population growth	Verhulst 1838
Branching processes, extinction of family names	Galton 1889
Correlation	Pearson 1903
Markov chains, statistics of language	Markov 1906
Hardy–Weinberg equilibrium in population genetics	Hardy 1908; Weinberg 1908
Analysis of variance, design of agricultural experiments	Fisher 1950
Dynamics of interacting species	Lotka 1925; Volterra 1931
Birth process, birth and death process	Yule 1925; Kendall 1948, 1949
Traveling waves in genetics	Fisher 1937; Kolmogorov et al. 1937
Game theory	von Neumann and Morgenstern 1953
Distribution for estimating bacterial mutation rates	Luria and Delbrück 1943
Morphogenesis	Turing 1952
Diffusion equation for gene frequencies	Kimura 1994
Circular interval graphs, genetic fine structure	Benzer 1959
Threshold functions of random graphs	Erdös and Rényi 1960
Sampling formula for haplotype frequencies	Ewens 1972
Coalescent genealogy of populations	Kingman 1982a, 1982b

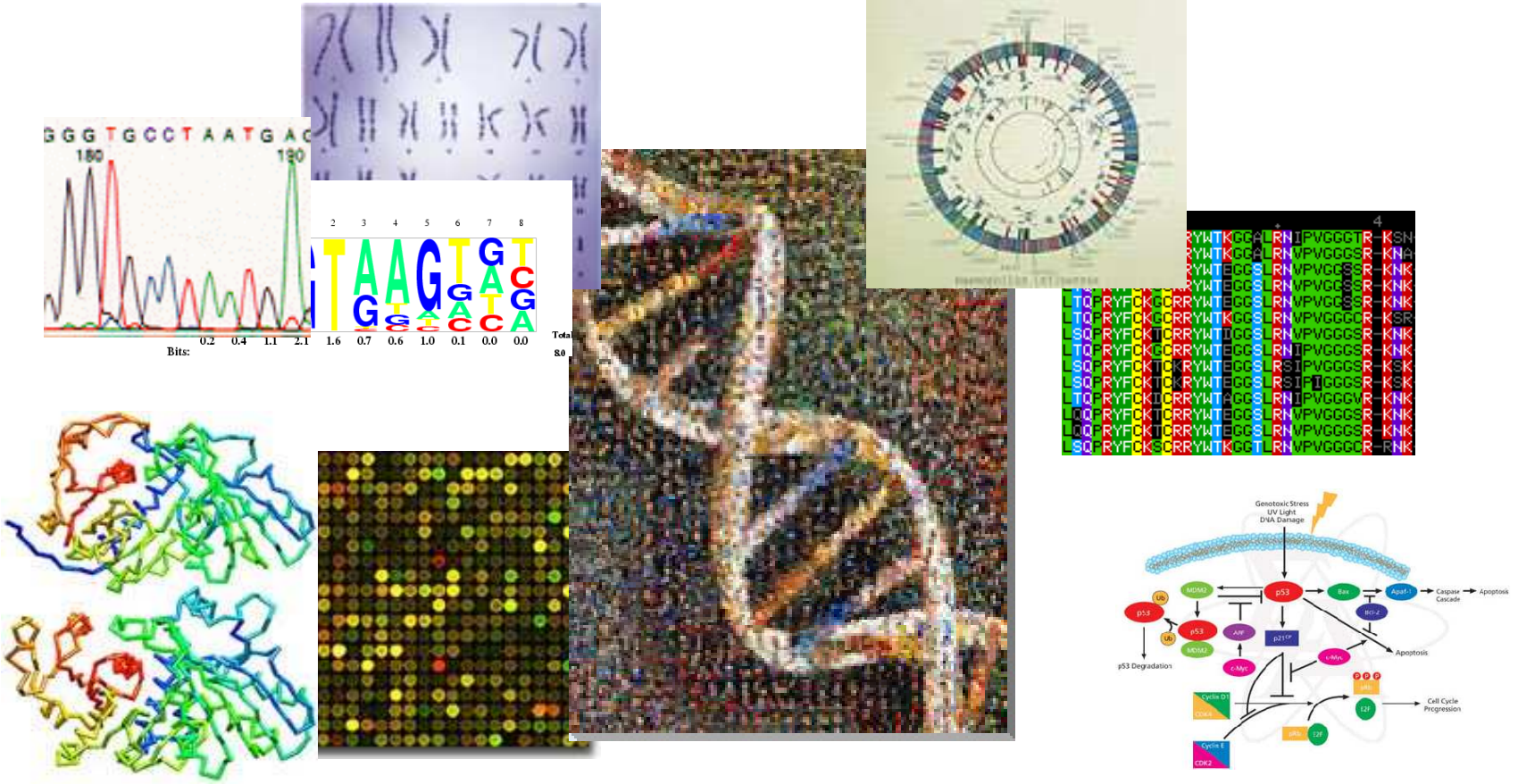
DOI: 10.1371/journal.pbio.0020439.t001

What is bioinformatics ?

The design, implementation, and use of computer algorithms to draw inferences from massive sets of biomolecular data.

An interdisciplinary science linking the life sciences to the computational, mathematical and statistical sciences.

What does bioinformatics involve?



PubMed

National Library of Medicine NLM

OMIM Online Mendelian Inheritance in Man — Johns Hopkins University

Modified from R.F. Yeh, UCSF

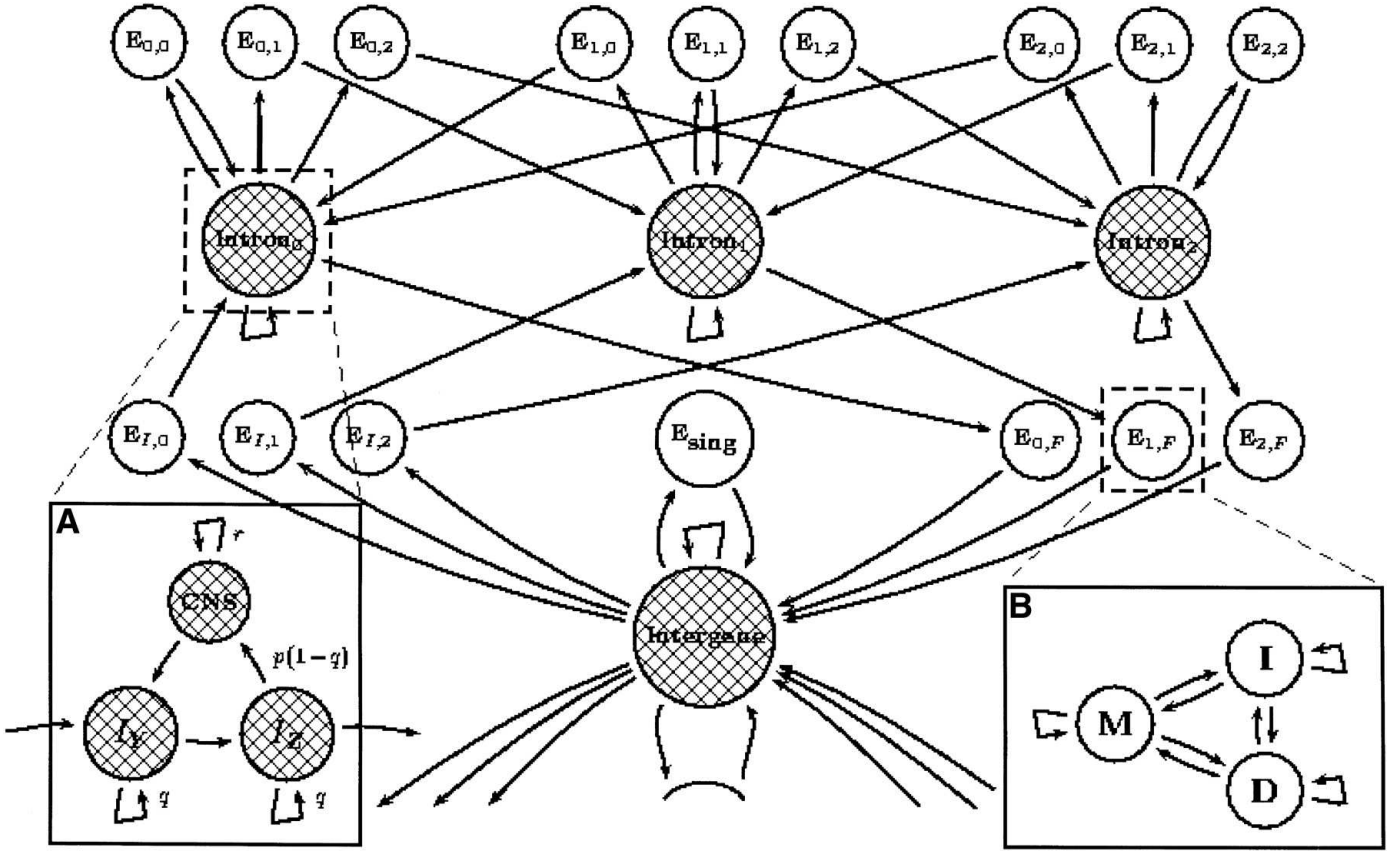
The emergence of bioinformatics

- With the discovery of DNA came a desire among biologist to understand how the genetic structure affects the biological traits of organisms.
- Sequencing is not sufficient (95% is junk DNA)
- Genes must be identify and their function need to be determine.
- Biological phenomena must link to protein identity, structure and genomic origin.

caaagcactaggattataggcgtgagccaccgtgcccggccaacaatgtcatttctttttcaggagtgggggtatccacagatgagctctggcttgtggggtccagtaagattggaggcgct
acttggggagggtgagccgctggtcccaggccaggcctggcctgtctccctcatgtgacttgaggaccaacagcacactgtggtttttgtctttatcagcaacaaatctgacacgctgcccctg
gccactcggtacaacgtcaagtgcgtggggctgtccccggatggccgcctcgctatcatcgtcgatgaaggtacttgcccttgatgtgggcgggtactgaggggaccagtgaggaggactca
gggctgtgtgggtctgaaatgatccgctccccagtcgcctccgtgttggggcccgggtggggattggggatgagggtgtcatgggcgaggctgcgtccccgccgcaggctgtgtctgtggcat
cacctgtggcagggcagggcctcatggcccttgggcattctcagtcctgactcctcacagtggacgcgagaaagacctttctccgtttttcccgtgctatgtatctggggtggcccctggggtagc
tgcgcggtggtgacctctggcgtcctgcaggggcgatgcgctgctggtcagcctggtctgcaggtctgtgctgcaccacttccacttcaagggctctgtgcacagtgtgtccttctcccctgatg
gcaggtaaggggacagcttgggggcaggagggttcggccgtgcatgccgacggtgaggcccactgcccagctcgtttccctgggccccgggttgtgacgttgggttgggcacacggccc
aagctctggcattggtgacccccttttctggaggagaaggcttggtgcggccaggctgccctccctccagggctgtttcttccctgacttttgcttcctctgctccttggcctccatcctgcaggagg
cgggactccagaaggcccaggtgggtgcaggctgagactccacagtggccttagggccattgctggtcttgaatatcaactatgtccgcgttccttccataggaagtttgttgtcacaaagggt
aacattgcccagatgtatcatgcccctgggaagaagcgggagttcaacgccttcgttctggacaagacctattttgggccctacgatgagaccacctgcatcgactggacggatgactccag
gtgcggcctcagaggcttcgggggagcggggcttgagagaggccccttggacaggtcacccccttgggccctgcagcttcgttccgcaatggcatctcggcctcctttggggctgggcgctg
cccacaccatcaggactggctgggtggctgtggtccagtctttgcggctctgaagtgtgtgctgtggccgagcctcctgcccctgcactgccctttcacctgtttgtcccagaaactgggaatgct
gtggttggctcgggcagcgtctgggaggagttaggaaggcagacagggcagggacagcgtctgggaggtgtgaggaaggcagacagggcaggggctgtgccggggtgtcttcctgtgc
ctggaggcgtctcagctcactctgagaccacgtggggctgtagcggtctccgtgcagccctctggagcttgcacctggctgctttgttatggccaagccctgaggaggtgccatctccctcttttc
caggtgctttgtggttgggagcaaagacatgtccacctgggtgttcggagccgagcgctgggacaacctcatctactatgcactgggggacataaggatgccatcgtggcctgcttctttgaa
tccaacagcctggacgtatgtccctttgccaagttccctagcctgatggtggccaaaagcggctagttcagggaggccccaggctgcaggcggcctcttcctgcggttcccccaggctgttccc
gtgtggcctgggggcatctagggaccgatggcagctgtggcggcccatcgagaattgttctcagagttcagaggggccgtcggtgccactggtccctggatctgagctggagggaggtgcc
gggttgtggcagctgggccagtgtggctggggctggggtgaccctgtgcttccccttgcagctgtactcactcagccaggacggagtgctgtgcatgtggcagtgtgacacgccccccgagg
gcttgcggctgaagccccctgcgggctggaaagcagacctgttgcagcgggaggaggaagaggaggaggaggaggaccaggagggcgacagagagaccaccatccggggaaaa
gccactccggccgaggaggagaagacaggaaaagtgaagtactcacggctggccaagtaggtctctgaggtgtggtgggctgtggcggggcactcctgtgggaatctcttgcccccaga
gggatctggatgtggggtccccagggccgaggagtgggaggtacagagagcaggcctgcgtctcagggccccggtttgagcaccctcgagtgcttgggtgttgaccctgccccttaagcag
ggtccctgcctcggctgtctgagctgcctgtgtgcccagggcagggcagagggcggtgggtaggggtgcataggggcagggaggtgtgcgggcactccgtgacgggaatagcagcac
cttgggccctaaggggttgggagttgtctgtggccccaccagcttctgtggagtttggtgtgggcgatagtcccttggcctcgtgcagggagatgccgggagagcatttggcaagggtggag
ggagggaccgggctgtgagtgggcgcagcgagtggtctgagtcctgcttcggggcccgctggtgggtggtacttggtgctttttcaggcactgagctgggccgtgcagccccgcatccggac
tcagtcccccagacgagacgagatgagatgagactgagcccggccccacggactgcctgtgggatggccttagctgtgcctggagccctggcagagatgctgccctgattcggggtgtccc
gcgtctgggctgtgggggaggctggtcaggatcacatgaaggcagagttgcccgcagctatggtgtctccgtcacccgggtgccccagcctgggcccctccttgctgcagttgcttagcgcc
tgctgtgtgcagggctctgaacttggcactgagacacactgcccacccttgtgaggcctcctctccgccggggaggagaggtgcagtgatgaggtggctgcgtgcggaggtgggcggcagt
tggggtccctgggtggggaaggcgcccacacaaggccctgcctgtggagggagccaggcggccagaacgggctgcctttggaggggttggggtctgctgggtaggtgggagttgtggcc
atgagagccctgggtctgaggagcagggagggccggctggactccagtgggaggagggtgctttgccaagggcagcggagttggagggtggccagaacaggtccaagtgggtcagtg
cccacaagtgagtcagtgcccgcgggtgggtcagtgcccgtgagtccgtgcccgtcagtgcccgcggggacttgggggtgttgtgctggagtcaagacttgaatggggactggaggaaga
ggaagggatgatgggccccccaaccccgccctggggaagggaggttccaggcagctccaggaagacagtgtgcctgctaggcggggactgcgggctgtcagcgcgaaacatgtgaag
ggagggcctgctgcattggcagccagtgctccctgtcacttgttcaggagggacggggtccagaaaggagagggaaggaccagggctgagctacaacgtgacttggggaggttttgagg
gcgggggtgccatcagggtcaccttcgtctgctgatggggaccgagggttccttagaggaggaagcagcaggggatgatgggattcgggtgtgcctggcggccatgccgcgcctcccgca
cctcctctgcctcccatgggcactgctgtctctgggtgcaccactagctctgctgtggggccagctggcgacagtggagatgctgtgtgaggttcctttggggtcccgagcttgggatggaggcg
ggcagcttcagggcctggggtggagagggagagggcaggtctgggacctcctgcctggatctgtggcctgtgtcctggagcccagagcaagctggggtggcagtctggctctgcgctcctc
ctcacaactcttctctctcgtcaattcaggtacttcttcaataaagaaggggattttaacaacctgacagctgcagcatttcataagaagtctcacctcttggtcactggctttgcttctggaatcttcc
atcttcatgagctgccagagtttaacctcatccactccctgaggtaagcctttgctcgcagtggggtgtggttttatgcactcactggccctgaatctgagggcccagccaagcctgcccttagca

caaagcactaggattataggcgtgagccaccgtgcccggccaacaatgtcatttctttttcaggagtgggggtatccacagatgagctctggcttgtggggtccagtaagattggaggcgct
acttggggagggtgagccgctggtcccaggccaggcctggcctgtctccctcatgtgacttgaggaccaacagcacactgtggtttttgtctttatcagcaacaaatctgacacgctgcccctg
gccactcggtacaacgtcaagtgcgtggggctgtccccggatggccgcctcgctatcatcgtcgatgaaggtacttgcccttgatgtgggcgggtactgaggggaccagtgaggaggactca
gggctgtgtgggtctgaaatgatccgctccccagtcgcctccgtgttggggcccgggtggggattggggatgagggtgtcatgggcgaggctgcgtccccgccgcaggctgtgtctgtggcat
cacctgtggcagggcagggcctcatggcccttgggcattctcagtcctgactcctcacagtggacgcgagaaagacctttctccgtttttcccgtgctatgtatctggggtggcccctggggtagc
tgcgcggtggtgacctctggcgtcctgcaggggcgatgcgctgctggtcagcctggtctgcaggtctgtgctgcaccacttccacttcaagggctctgtgcacagtgtgtccttctcccctgatg
gcaggtaaggggacagcttgggggcaggagggttcggccgtgcatgccgacggtgaggcccactgcccagctcgtttccctgggccccgggttgtgacgttgggttgggcacacggccc
aagctctggcattggtgacccccttttctggaggagaaggcttggtgcggccaggctgccctccctccagggctgtttcttccctgacttttgcttcctctgctccttggcctccatcctgcaggagg
cgggactccagaaggcccaggtgggtgcaggctgagactccacagtggccttagggccattgctggtcttgaatatcaactatgtccgcgttccttccataggaagtttgttgtcacaaagggt
aacattgcccagatgtatcatgcccctgggaagaagcgggagttcaacgccttcgttctggacaagacctattttgggccctacgatgagaccacctgcatcgactggacggatgactccag
gtgcggcctcagaggcttcgggggagcggggcttgagagaggccccttggacaggtcacccccttgggccctgcagcttcgttccgcaatggcatctcggcctcctttggggctgggcgctg
cccacaccatcaggactggctgggtggctgtggtccagtctttgcggctctgaagtgtgtgctgtggccgagcctcctgcccctgcactgccctttcacctgtttgtcccagaaactgggaatgct
gtggttggctcgggcagcgtctgggaggagttaggaaggcagacagggcagggacagcgtctgggaggtgtgaggaaggcagacagggcaggggctgtgccggggtgtcttcctgtgc
ctggaggcgtctcagctcactctgagaccacgtggggctgtagcggtctccgtgcagccctctggagcttgcacctggctgctttgttatggccaagccctgaggaggtgccatctccctcttttc
caggtgctttgtggttgggagcaaagacatgtccacctgggtgttcggagccgagcgctgggacaacctcatctactatgcactgggggacataaggatgccatcgtggcctgcttctttgaa
tccaacagcctggacgtatgtccctttgccaagttccctagcctgatggtggccaaaagcggctagttcagggaggccccaggctgcaggcggcctcttcctgcggttcccccaggctgttccc
gtgtggcctgggggcatctagggaccgatggcagctgtggcggcccatcgagaattgttctcagagttcagaggggccgtcggtgccactggtccctggatctgagctggagggaggtgcc
gggttgtggcagctgggccagtgtggctggggctggggtgaccctgtgcttccccttgcagctgtactcactcagccaggacggagtgctgtgcatgtggcagtgtgacacgccccccgagg
gcttgcggctgaagccccctgcgggctggaaagcagacctgttgcagcgggaggaggaagaggaggaggaggaggaccaggagggcgacagagagaccaccatccggggaaaa
gccactccggccgaggaggagaagacaggaaaagtgaagtactcacggctggccaagtaggtctctgaggtgtggtgggctgtggcggggcactcctgtgggaatctcttgcccccaga
gggatctggatgtggggtccccagggccgaggagtgggaggtacagagagcaggcctgcgtctcagggccccggtttgagcaccctcgagtgcttgggtgttgaccctgccccttaagcag
ggtccctgcctcggctgtctgagctgcctgtgtgcccagggcagggcagagggcggtgggtaggggtgcataggggcagggaggtgtgcgggcactccgtgacgggaatagcagcac
cttgggccctaaggggttgggagttgtctgtggccccaccagcttctgtggagtttggtgtgggcgatagtcccttggcctcgtgcagggagatgccgggagagcatttggcaagggtggag
ggagggaccgggctgtgagtgggcgcagcgagtggtctgagtcctgcttcggggcccgctggtgggtggtacttggtgctttttcaggcactgagctgggccgtgcagccccgcatccggac
tcagtcccccagacgagacgagatgagatgagactgagcccggccccacggactgcctgtgggatggccttagctgtgcctggagccctggcagagatgctgccctgattcggggtgtccc
gcgtctgggctgtgggggaggctggtcaggatcacatgaaggcagagttgcccgcagctatggtgtctccgtcacccgggtgccccagcctgggcccctccttgctgcagttgcttagcgcc
tgctgtgtgcagggctctgaacttggcactgagacacactgcccacccttgtgaggcctcctctccgccggggaggagaggtgcagtgatgaggtggctgcgtgcggaggtgggcggcagt
tggggtccctgggtggggaaggcgcccacacaaggccctgcctgtggagggagccaggcggccagaacgggctgcctttggaggggttggggtctgctgggtaggtgggagttgtggcc
atgagagccctgggtctgaggagcagggagggccggctggactccagtgggaggagggtgctttgccaagggcagcggagttggagggtggccagaacaggtccaagtgggtcagtg
cccacaagtgagtcagtgcccgcgggtgggtcagtgcccgtgagtccgtgcccgtcagtgcccgcggggacttgggggtgttgtgctggagtcaagacttgaatggggactggaggaaga
ggaagggatgatgggccccccaaccccgccctggggaagggaggttccaggcagctccaggaagacagtgtgcctgctaggcggggactgcgggctgtcagcgcgaaacatgtgaag
ggagggcctgctgcattggcagccagtgctccctgtcacttgttcaggagggacggggtccagaaaggagagggaaggaccagggctgagctacaacgtgacttggggaggttttgagg
gcgggggtgccatcagggtcaccttcgtctgctgatggggaccgagggttccttagaggaggaagcagcaggggatgatgggattcgggtgtgcctggcggccatgccgcgcctcccgca
cctcctctgcctcccatgggcactgctgtctctgggtgcaccactagctctgctgtggggccagctggcgacagtggagatgctgtgtgaggttcctttggggtcccgagcttgggatggaggcg
ggcagcttcagggcctggggtggagagggagagggcaggtctgggacctcctgcctggatctgtggcctgtgtcctggagcccagagcaagctggggtggcagtctggctctgcgctcctc
ctcacaactcttctctctcgtcaattcaggtacttcttcaataaagaaggggattttaacaacctgacagctgcagcatttcataagaagtctcacctcttggtcactggctttgcttctggaatcttcc
atcttcatgagctgccagagtttaacctcatccactccctgaggtaagcctttgctcgcagtggggtgtggttttatgcactcactggccctgaatctgagggcccagccaagcctgcccttagca

Generalized Pair HMM for simultaneous human-mouse gene predictions (SLAM)



Alexandersson et al. Genome Res. 2003; 13: 496-502

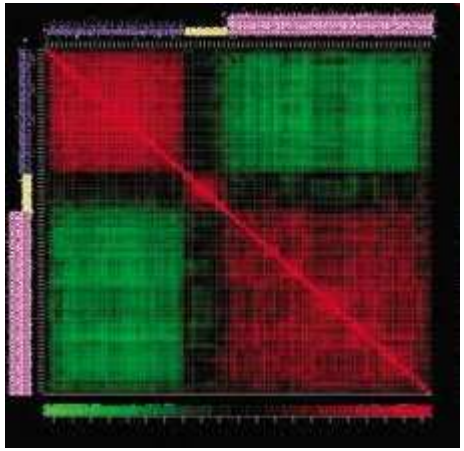
Post-genome biology

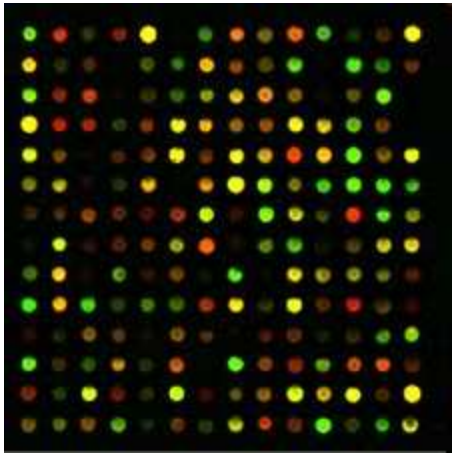
- What does each gene do?
- How are the genes regulated?
- What are the roles of genes in each disease?
  - Disease processes
  - Molecular diagnosis
  - Genetic predisposition

Ultimately,

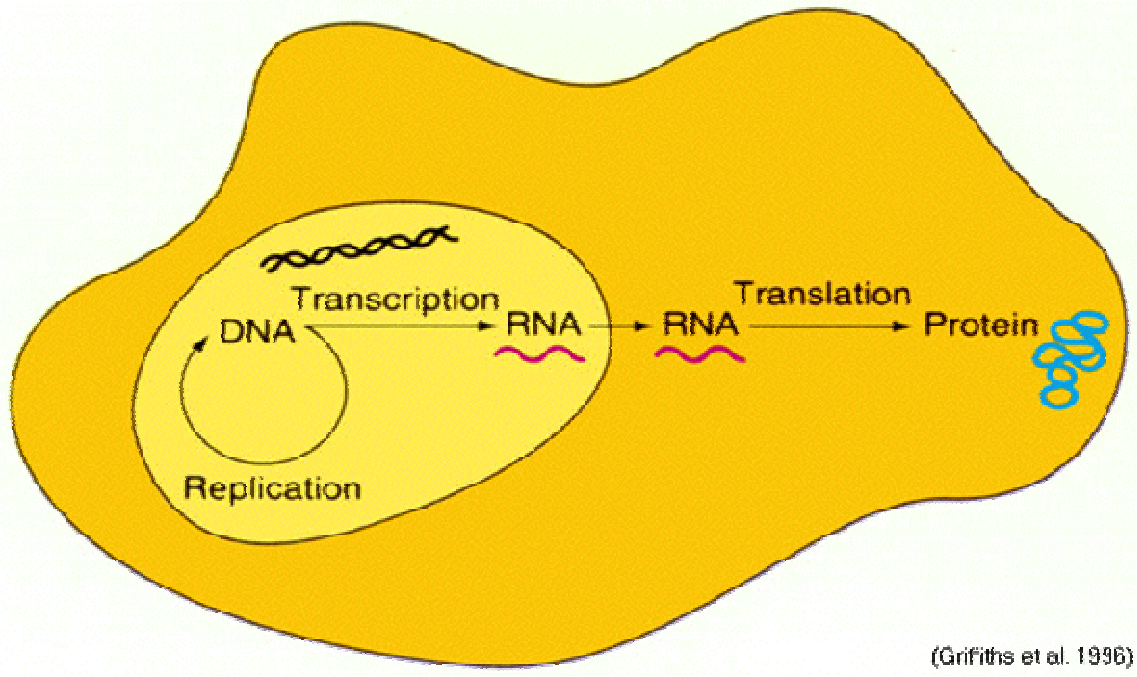
to understand life and provide personalized medicine

Gene expression - microarrays





Gene expression



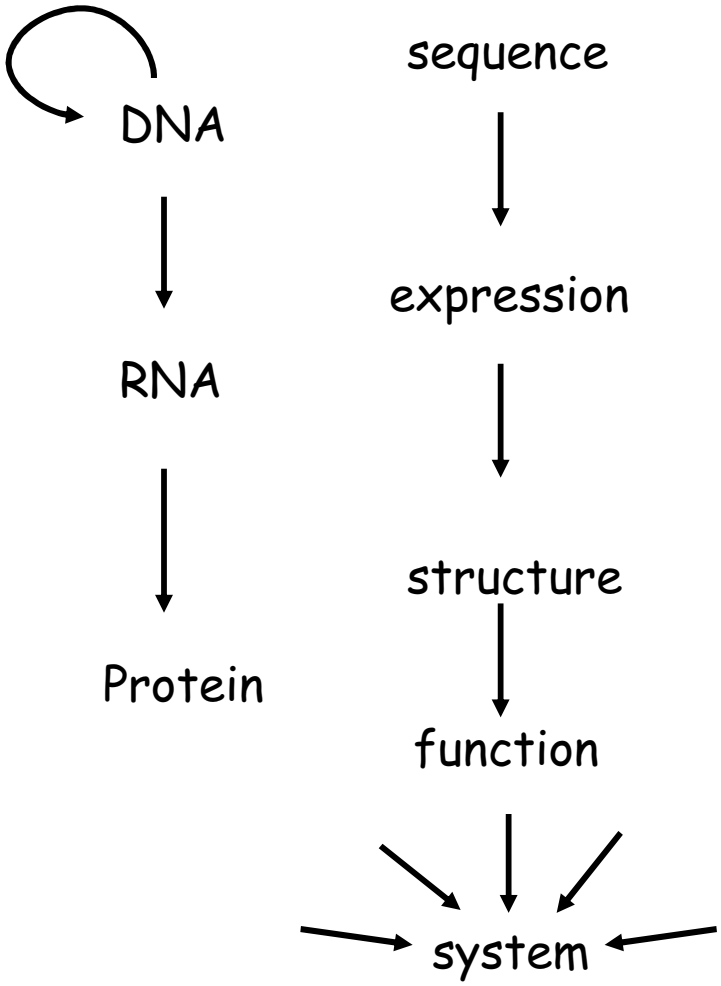
The **expression** of the genetic information stored in the DNA molecule occurs in two stages:

(i) **transcription**, during which DNA is transcribed into mRNA;

(ii) **translation**, during which mRNA is translated to produce a protein.

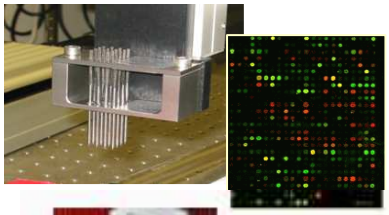
Idea: measure the amount of mRNA to see which genes are being expressed in (used by) the cell.

Biology as Information Science: in Large Scale!

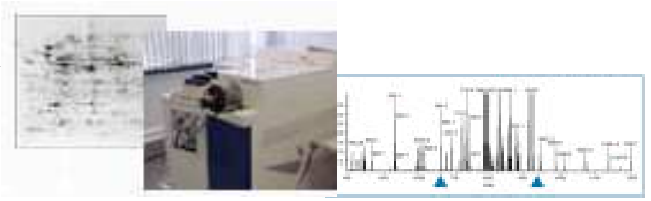


Genome

Transcriptome

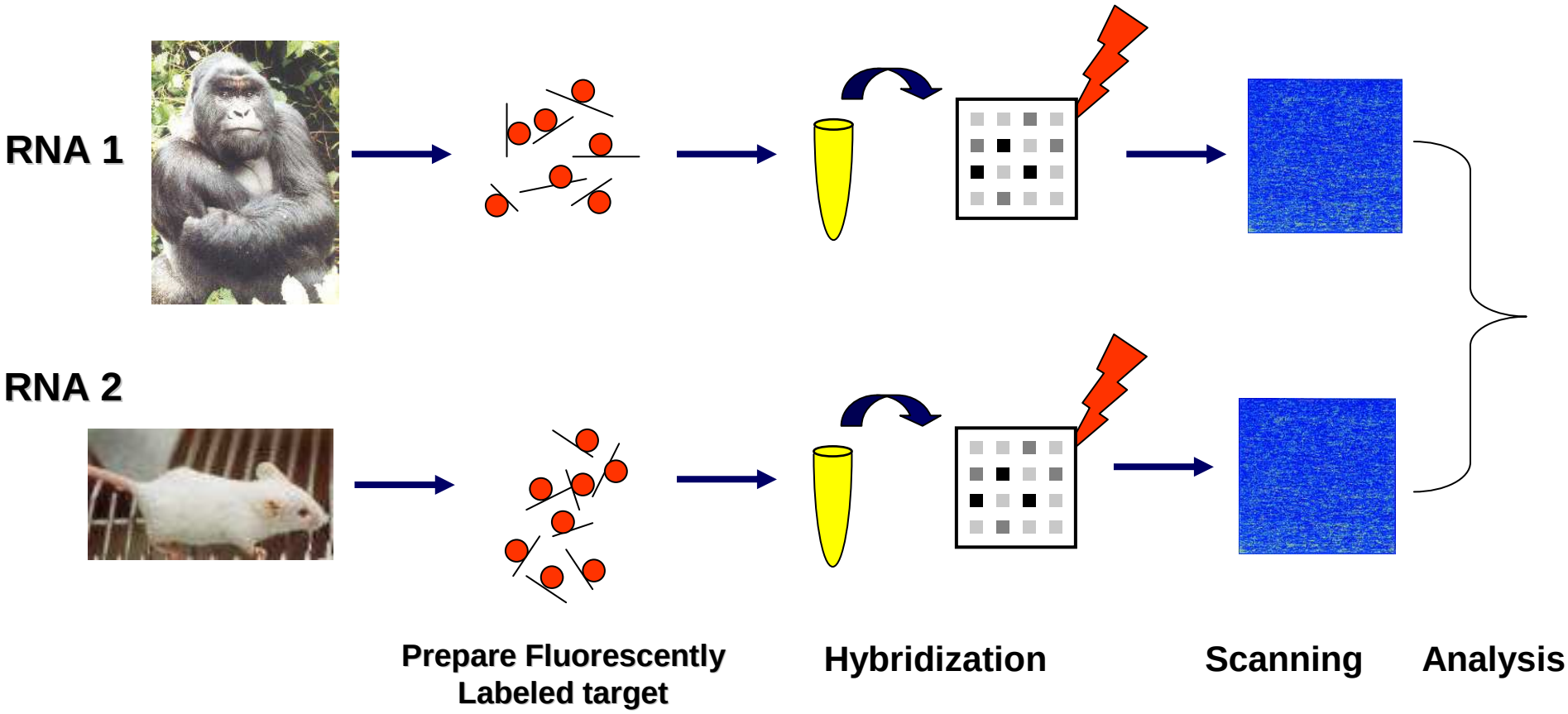


Proteome



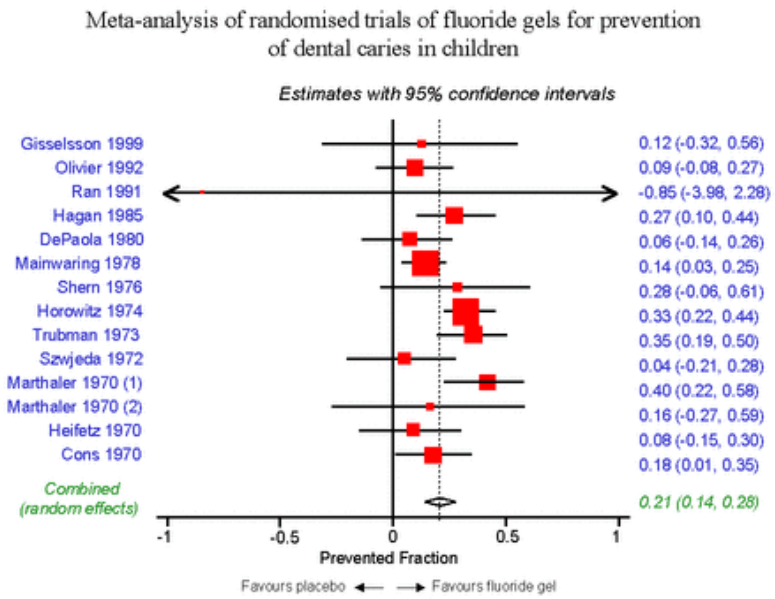
Systems biology
(Regulome, Epigenome, Metabolome,Physiome...)

Expression profiling with Affymetrix arrays



Meta-analysis in biostatistics

- Meta-analysis involves combining summary information from related but independent studies.
- Increasing power to detect an overall treatment effect.
- Assessment of the amount of variability between studies.
- Publication bias



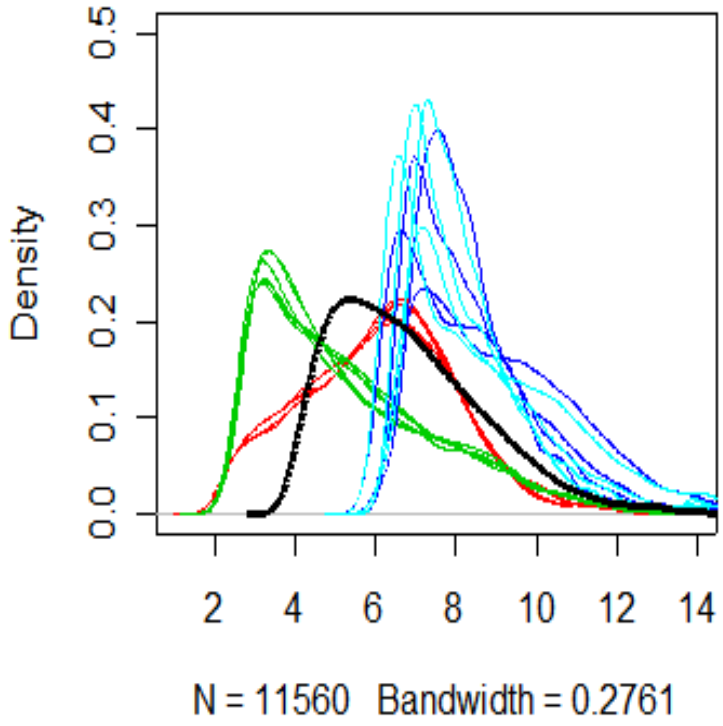
Identify DE genes

- Looking for concordance; transformation to a unit-less metric (statistics)
  - Correlation.
  - T or F-statistics or p-values

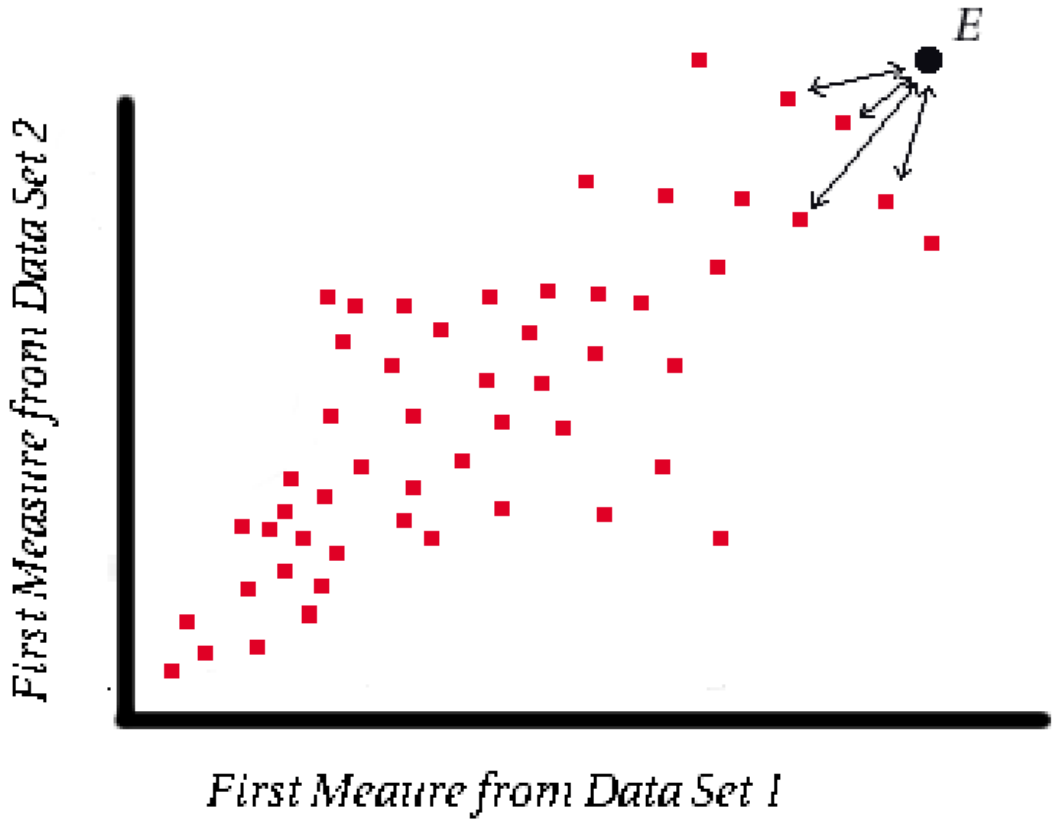
  Example:

  $S = -2^* \log P_1 + -2^* \log P_2 + \ldots + -2^* \log P_n$

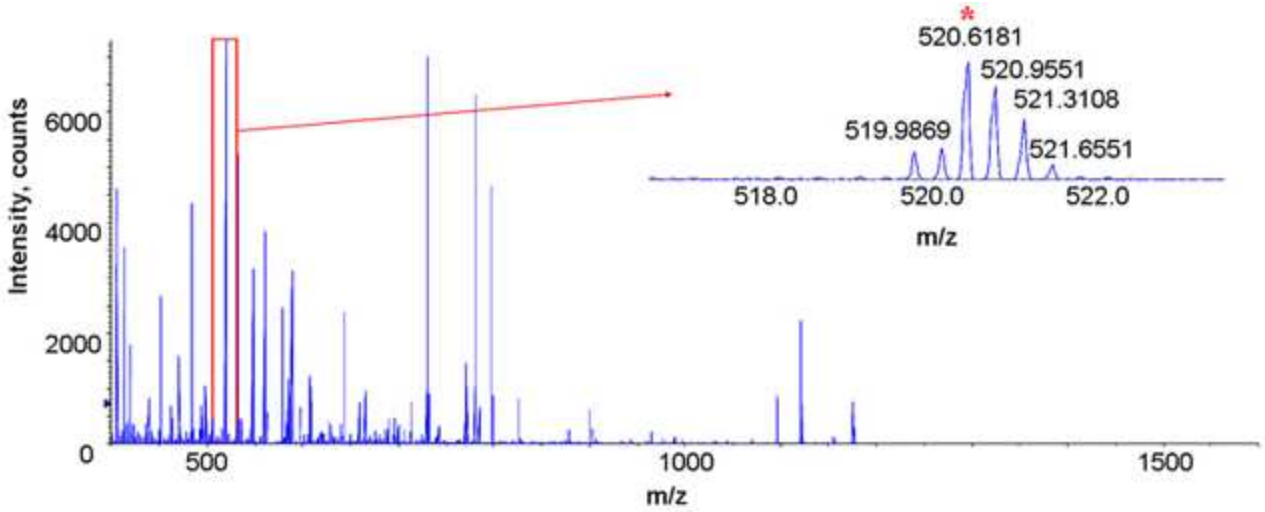
- Comparing genes quantitatively.
  - Normalization.
  - Removing unwanted variation.
  - Linear modeling with platform as
  - an effect.



mDEDS measure

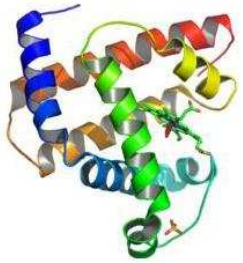


Protein expression - iTRAQ



Proteins

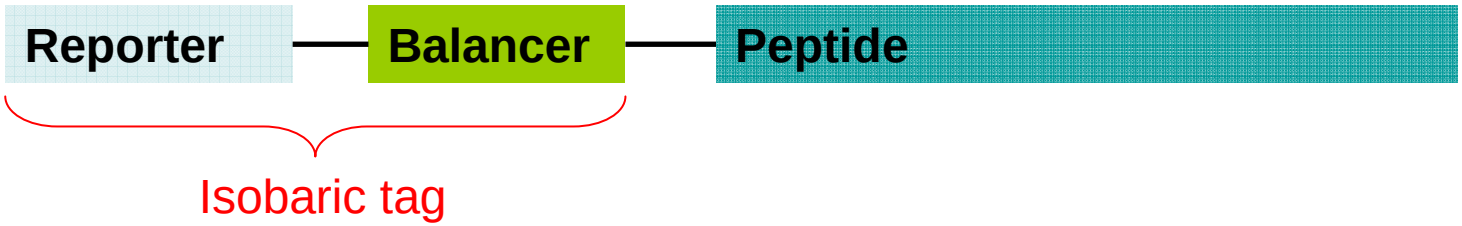
A L D L K P G Q S



Peptide sequence

iTRAQ : isobaric tag for relative and absolute quantitation

- Identify and quantify proteins through MS/MS.
- Analyses up to 8 samples simultaneously
- isobaric tag labelling
  - Reporter (m/z: 112-119) + balancer



  - Gives identical peptides from different samples the same mass

Ross et al, *Mol. Cell. Proteomics* **3**, 2004

Comparing protein expression

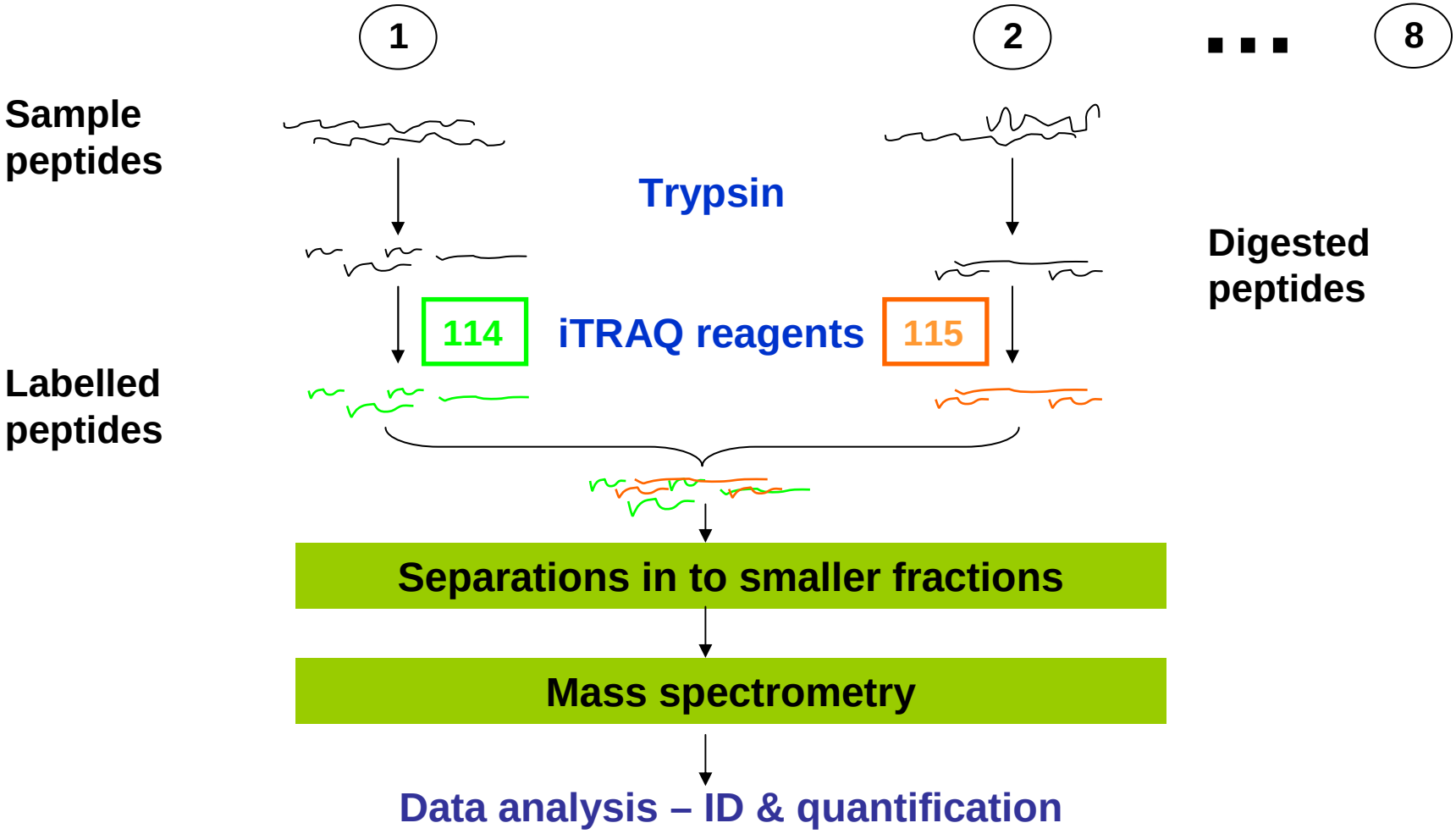
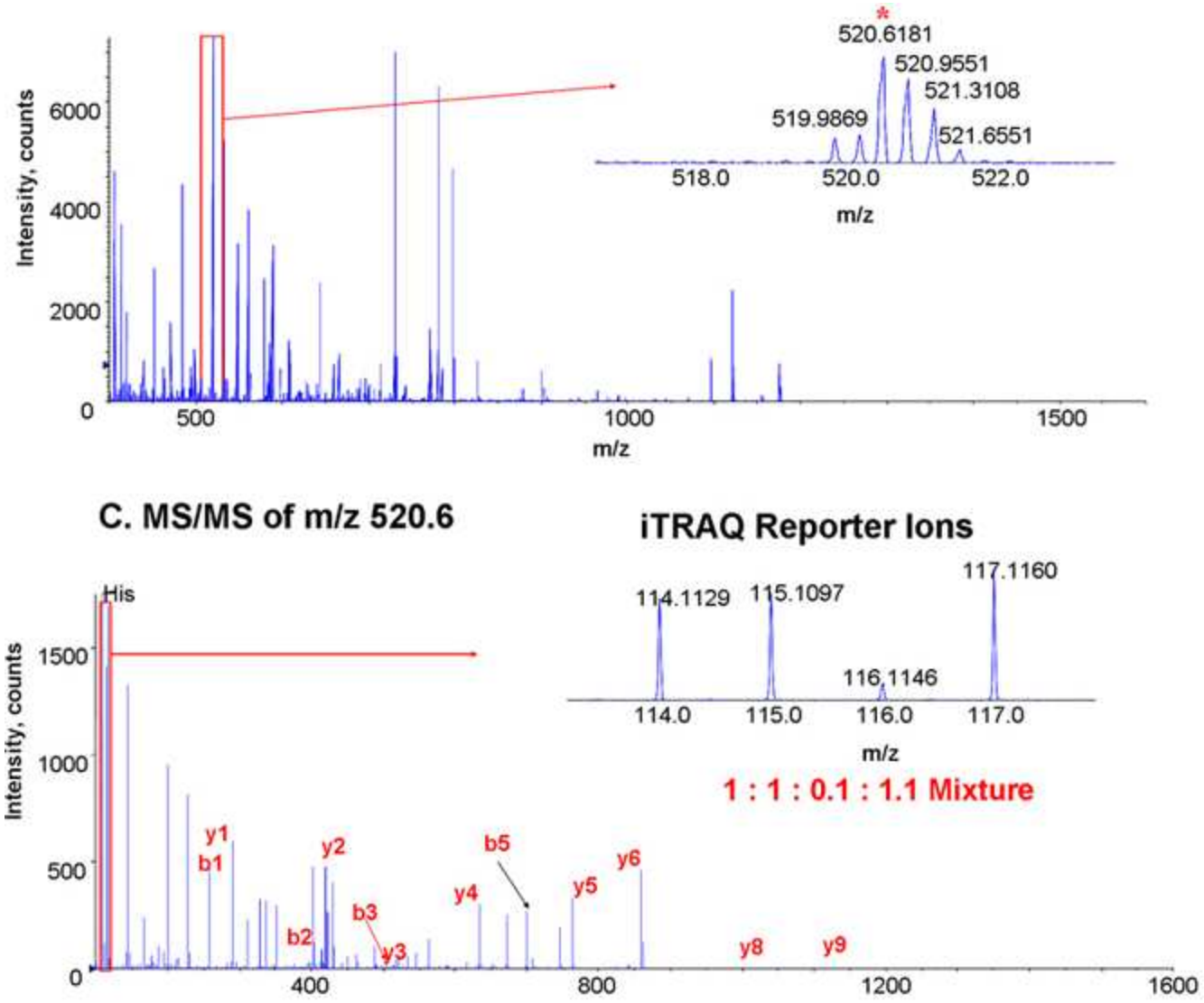
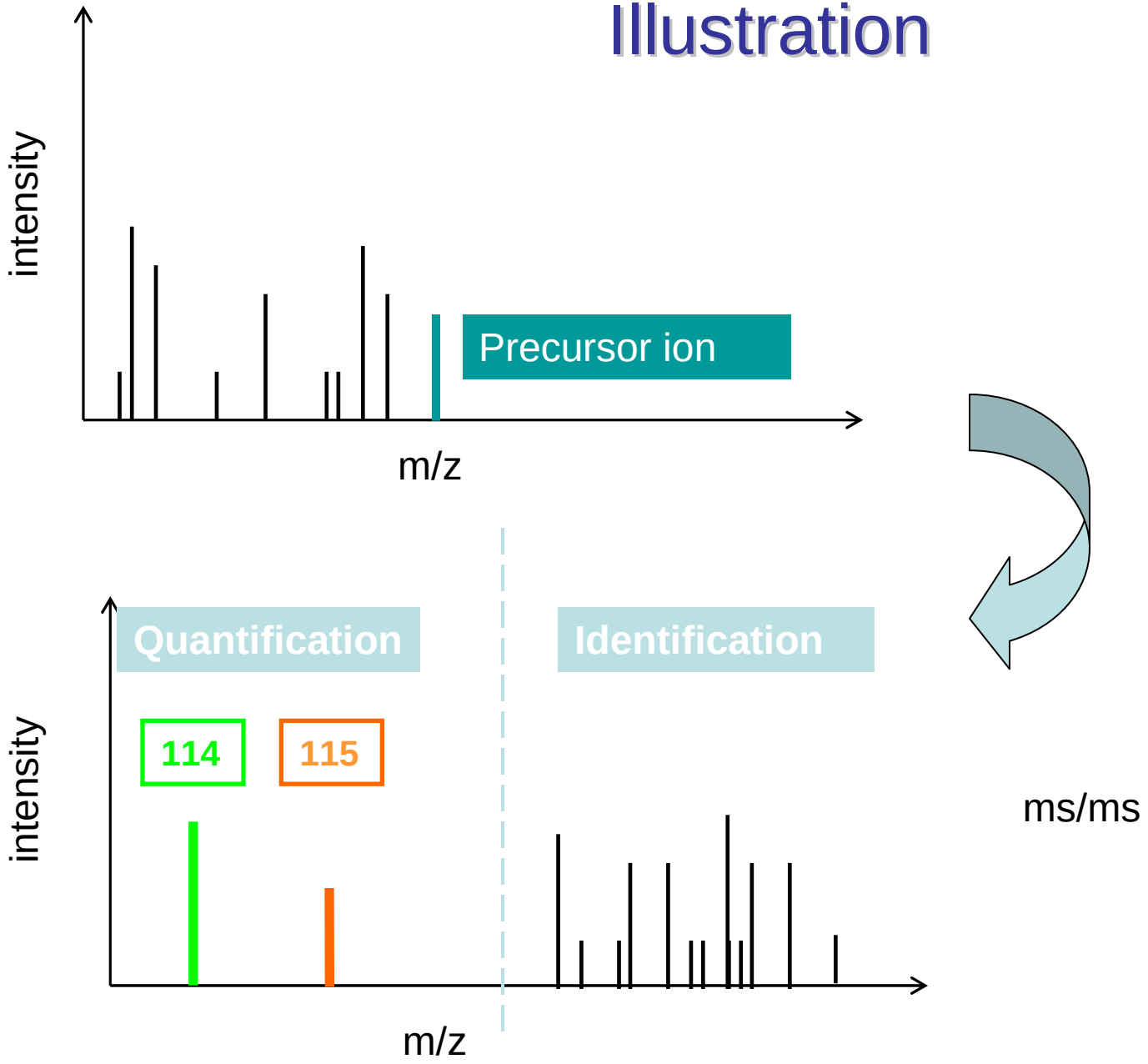


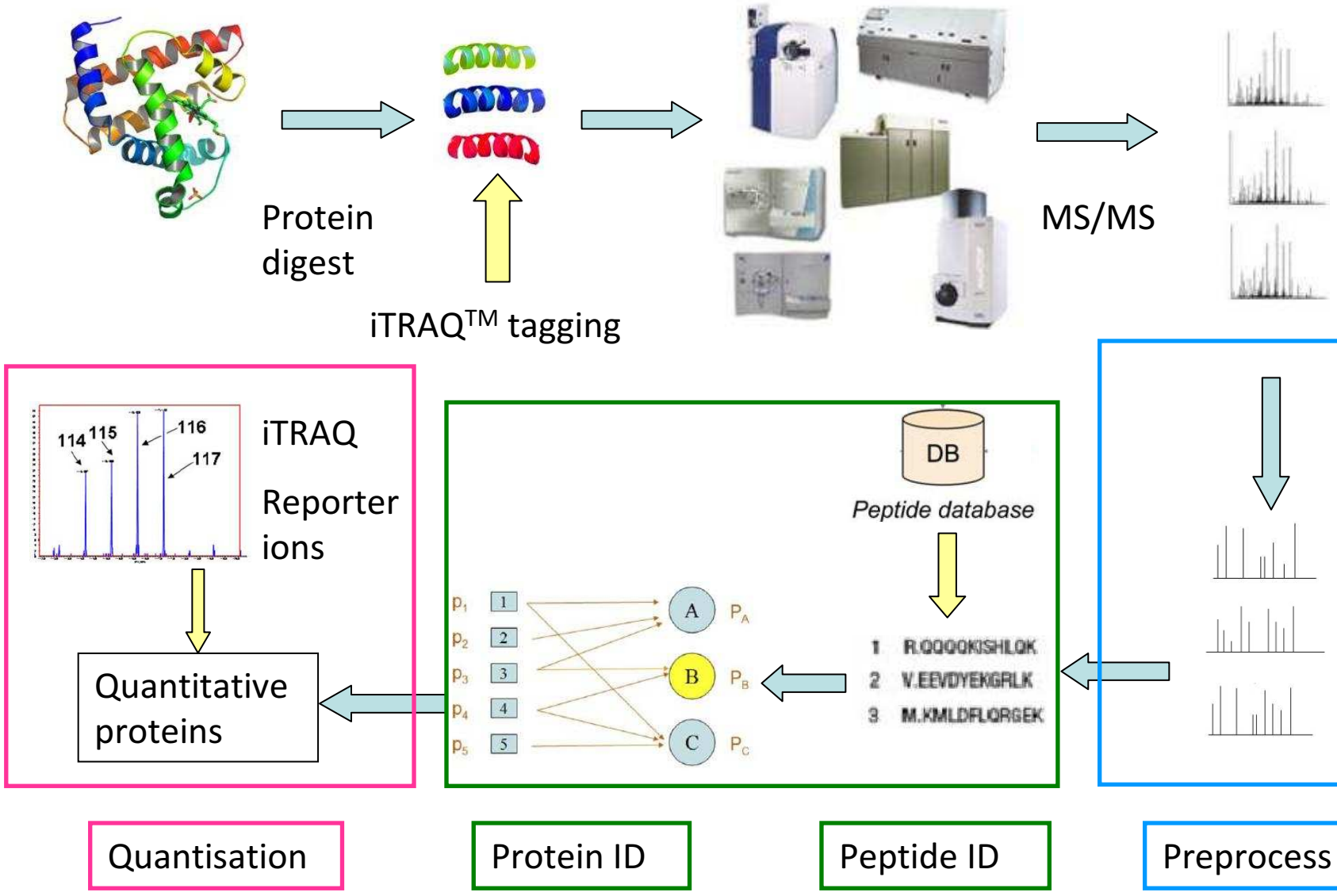
Illustration: ms/ms



Illustration



iTRAQ data analysis



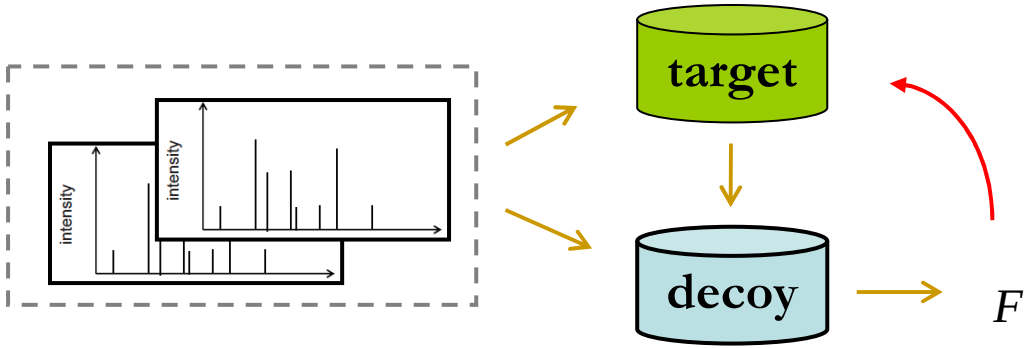
An example of processed iTRAQ experiment ….

		Control	Trt 1	Trt 2	Trt 2	Trt1	Trt 3	Trt 3	Empty
Protein	**Peptide**	**114.1**	**115.1**	**116.1**	**117.1**	**114.1**	**115.1**	**116.1**	**117.1**
sp|Q9NVF7	TNGAGVTVLR	448	96	183	6282	93	4550	3750	12
	EEGGGGQGD	14	0	0	187				
sp|Q15751	DLSQPESDEE					99	17	128	0
sp|Q9NRC6	RLQTEACRLGQLHPAA	365	39	0	211	450	452	395	0
sp|A6BM72	GCQLPCQCRHGASCDP	0	0	0	60				
	CPSGTWGLNCNESCTC	23	0	0	204				
sp|Q01955|	PGRQGAAGLKG	43	16	47	21	18	200	149	12
	LLPLLLVLLAAA	212	18	223	0	20	223	150	0
	GLPGFSGSPGLPGTP	166	93	282	122	120	231	149	0

- Experimental design … how many samples?
- How to model missing values?

Target decoy search method

- Create a decoy protein database from the target database and search the input spectra upon the decoy database. Each decoy (peptide)-spectrum match is treated as a false discovery.
- The overall decoy-spectrum matches is used to estimate the FDR for the target database.

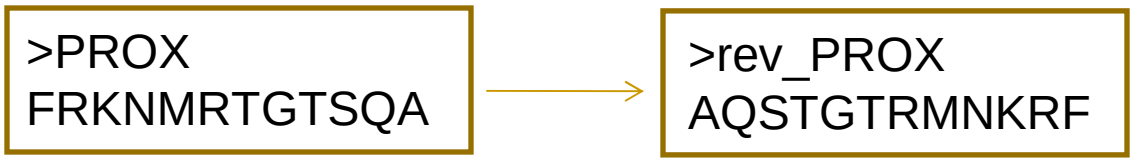


Assumptions:

- Target and decoy databases do not overlap.
- Target and decoy false positives are equally likely to be selected by the search algorithm.

Elias & Gygi, *Nature Methods* - 4, 207-214 (2007).
Gyqi et al., *Nature Methods*, Vol. 2, - 9., 667-675 (2005).

Reverse based approach



Features:

(1) Produce a decoy database with equal size to the target database

(2) Create a decoy sequence for each protein entry individually

(3) Completely preserve the general composition of the target database

Can reverse based decoy database faithfully estimate false positive identifications (FPI) in target database?

```
Q05639 FIKNMITGTSQADCAVLIVAAGVGEFEAGISKNGQTREHALLAYTLGVKQ
P68104 FIKNMITGTSQADCAVLIVAAGVGEFEAGISKNGQTREHALLAYTLGVKQ
Q5VTE0 FIKNMITGTSQADCAVLIVAAGVGEFEAGISKNGQTREHALLAYTLGVKQ
Q9Y450 FIPNMITGAAQADVAVLVVDASRGEFEAGFETGGQTREHGLLVRSLGVTQ
P15170 FVPNMIGGASQADLAVLVISARKGEFETGFEKGGQTREHAMLAKTAGVKH
       *: *** *:::** ***:: *  ****:*:...******.:*. : **.:
```

Data

- Seattle spectra dataset is a protein mixture analysed by the ABI 4700 spectrometer. The dataset contains 18 known proteins derived from multiple organisms (Bovine, *E. coli*, *B. licheniformis*, Rabbit, Horse, and Chicken)
  - Target database used: SWISS-PROT sequence library
- Aurum spectra dataset is generated by MALDI TOF/TOF. It contains 246 known proteins from human-only protein mixture with 100 possible contaminants.
  - Target database used: human specific SWISS-PROT sequence library

Dataset	# Proteins	# Contaminations	Organism
Seattle	18	15	Multiple Organisms
Aurum	246	100	Homo Sapiens

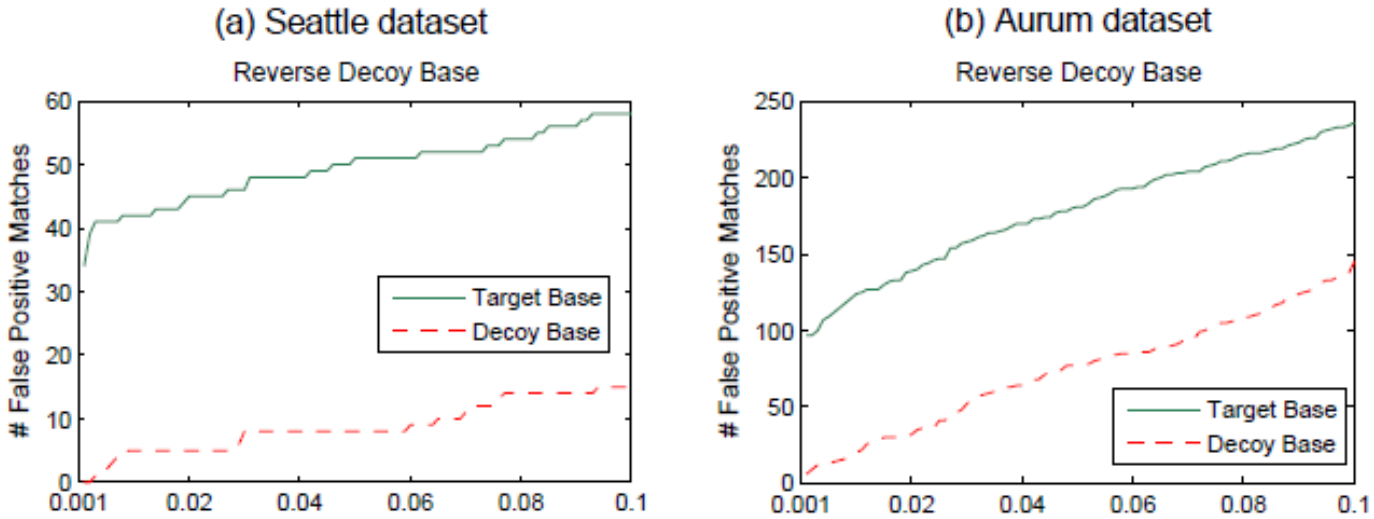
J.A. Falkner et al. *Journal of the American Society for Mass Spectrometry*, 18(5):850–855, 2007.

J. Klimek et al *Journal of Proteome Research*, 7(1):96–103, 2008.

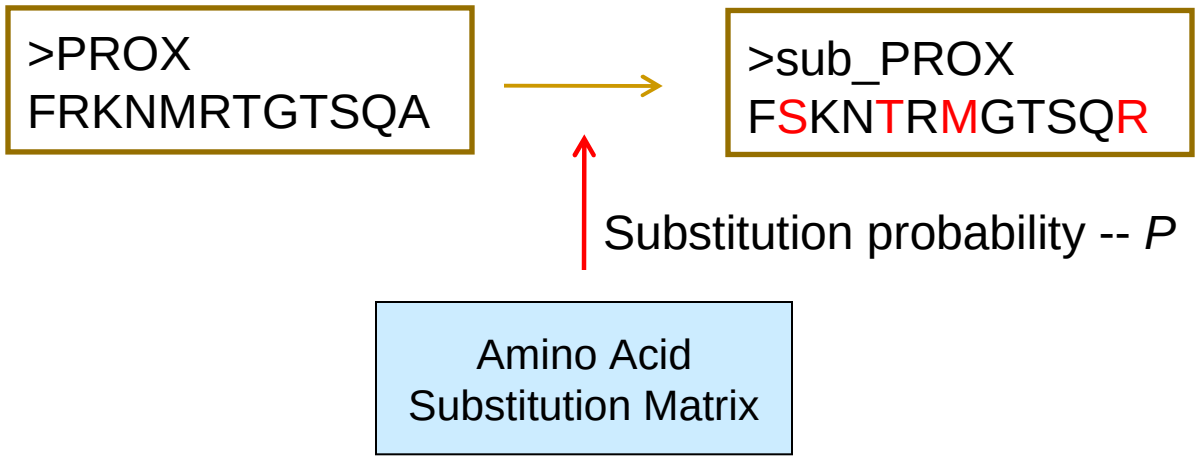
Simulation

Examine equal likelihood of false positive identification (FPI).

- Removing the known protein sequences of the reference dataset from the target database.
- Concatenating the target database and the decoy database and search the reference spectra against the concatenated database.
- Comparing the number of matches from the target database and the decoy database.



Substitution based approach



Algorithm 1 Substitution based decoy base construction

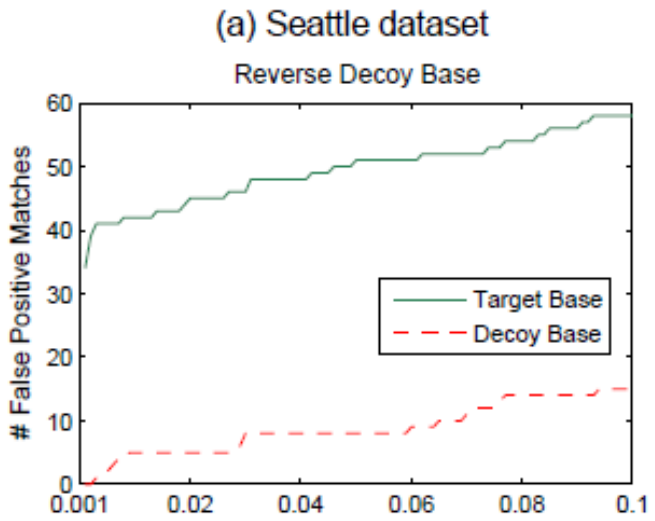
```
1: Input: target base; p {probability}
2: Output: decoy base
3: for each protein sequence do
4:    for each amino acid do
5:       if p fulfilled then
6:          substitute current amino acid using substitution matrix;
7:       end if
8:    end for
9: end for
```

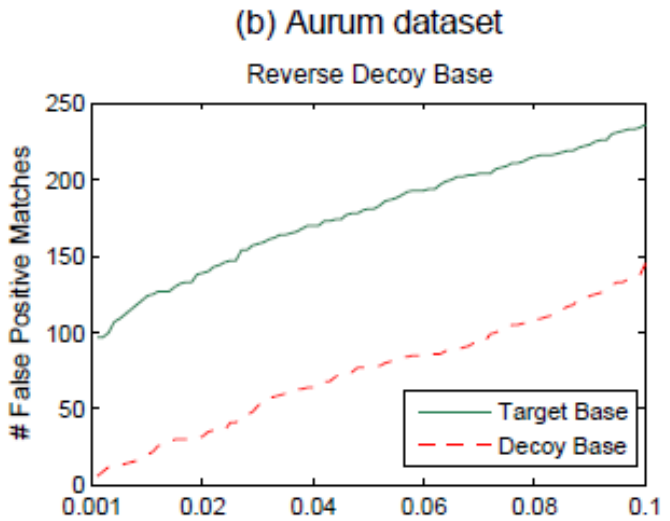
Two key parameters:

1. Substitution probability value p.
2. Substitution matrix M.

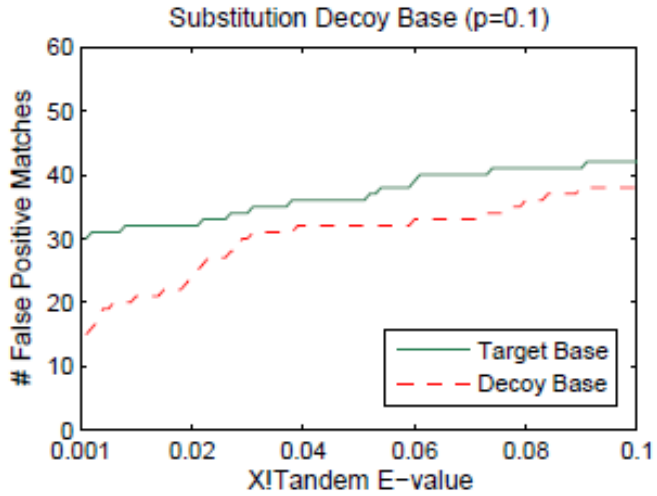
Comparison with reverse decoy approach

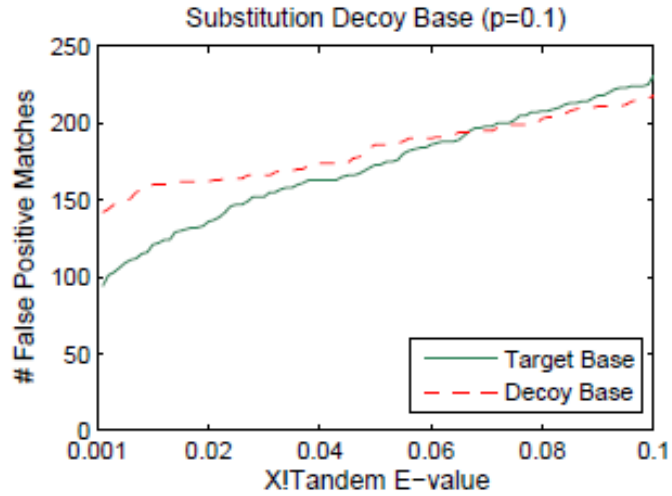
Reverse





Substitution





Other current problems …

- Sequence alignment
- Gene finding
- Genome assembly
- Protein structure alignment
- Prediction of gene expression
- Protein-protein interaction
- Modeling evolution

Acknowledgments

School of Mathematics and statistics

- Anna Campain
- Ellis Patrick
- Penghao Wang
- Pengyi Yang
- Vivek Jayaswal

Central Clinical School, University of Sydney

– Dr. Jonathan Arthur

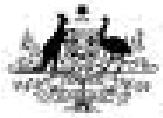
Sydney University Proteome Research Unit (SUPRU)

– Dr. Ben Crossett

School of Molecular & Microbial Biosciences, University of Sydney

– Philippa Kohnke

– Prof. Richard Christopherson



Australian Government

Australian Research Council