CMSC423: Bioinformatic Algorithms, Databases and Tools

What you missed

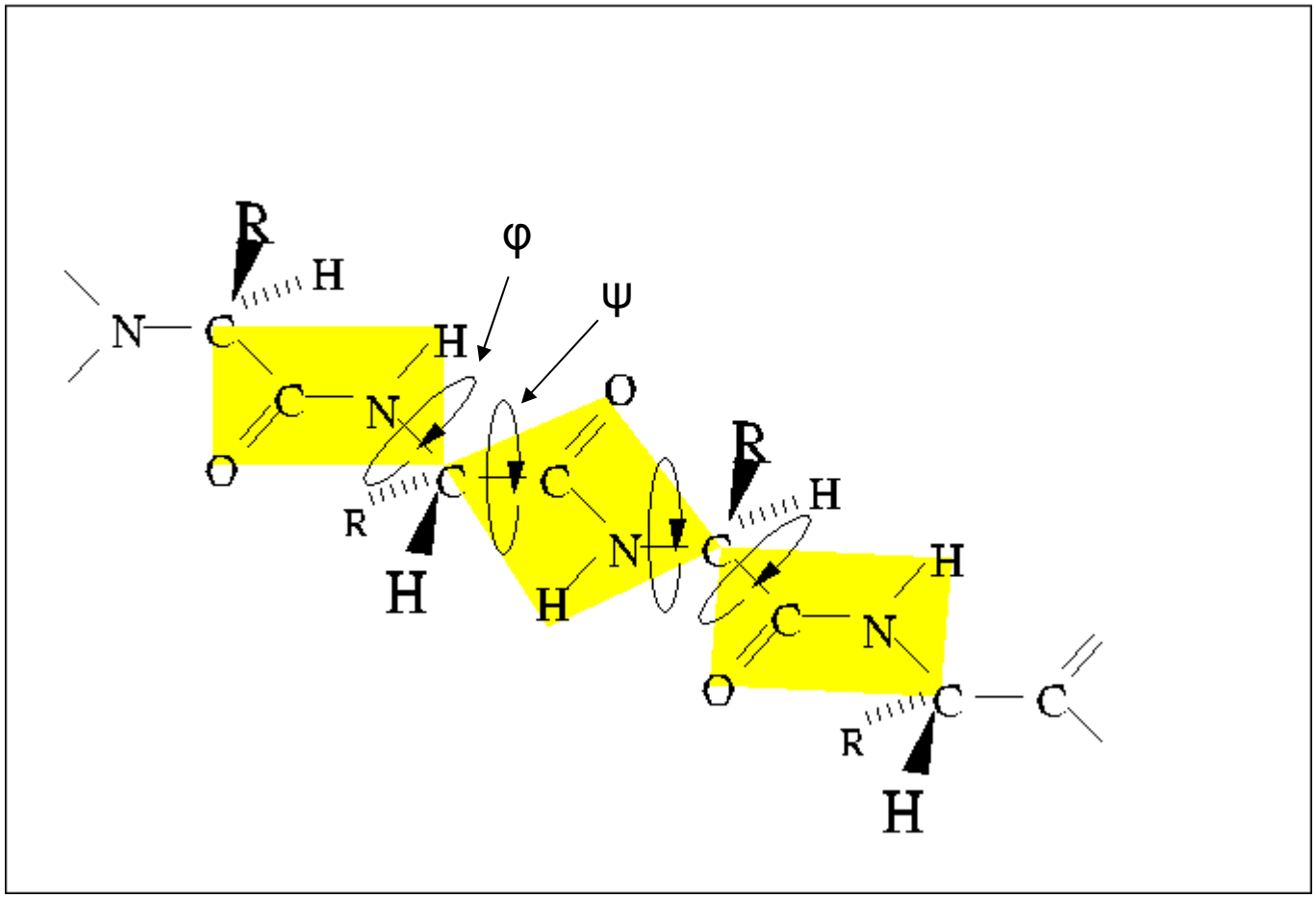
Protein folding

Protein folding

- Note: mis-folded proteins may cause disease (e.g. Creutzfeld-Jakob a.k.a. mad cow)
- Drugs (e.g. antibiotics) often inhibit protein function – knowing structure can help design drugs

- Folding@home – lend your computer's unused cycles to help fold proteins (like SETI@home) (do you believe in evolution or aliens ?)

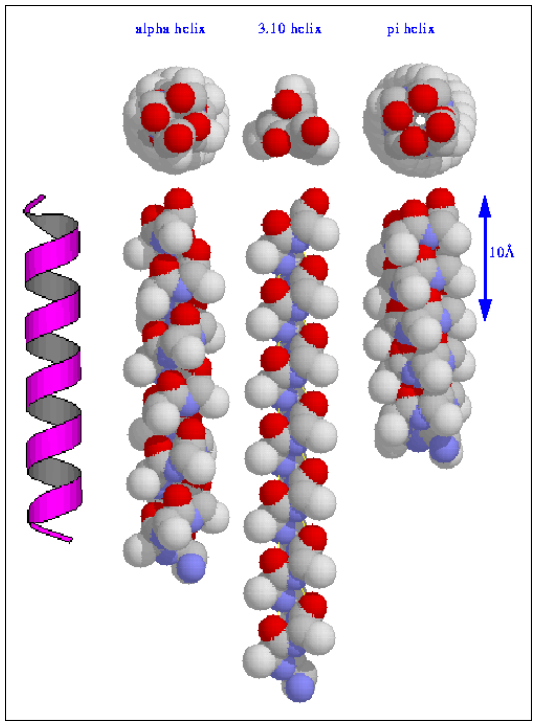
Protein structure
(primary structure = sequence)



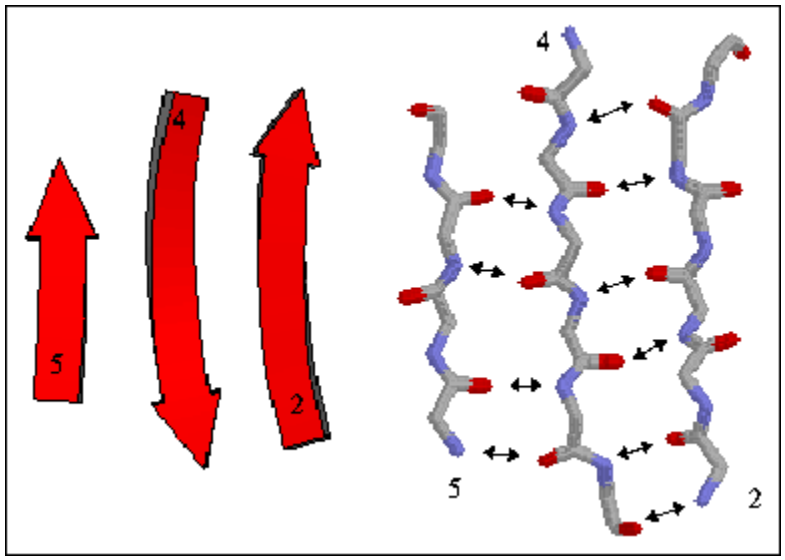
http://www.tulane.edu/~biochem/med/second.htm

Secondary structure (motifs)

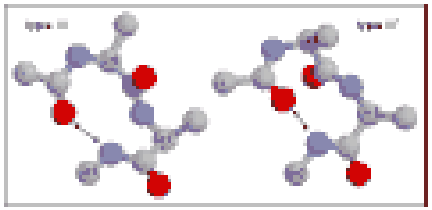
helix



sheet

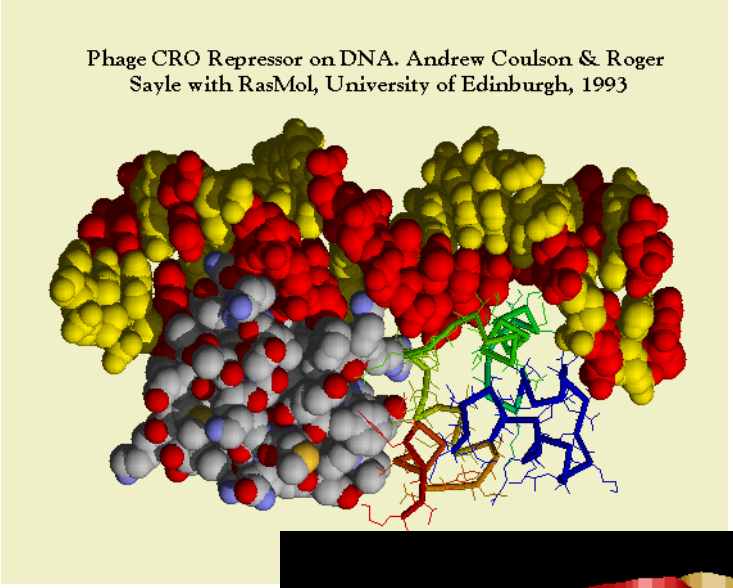


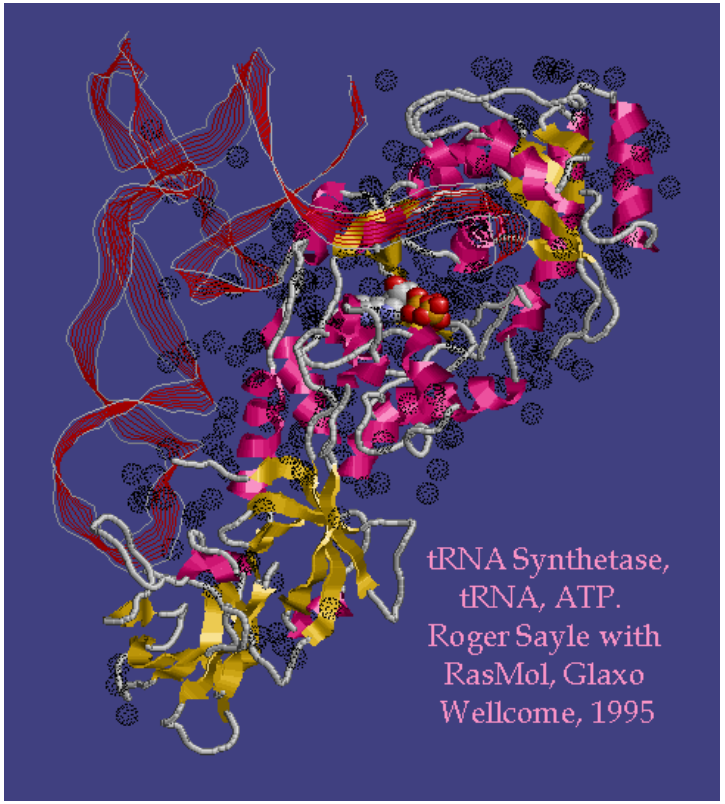
turn

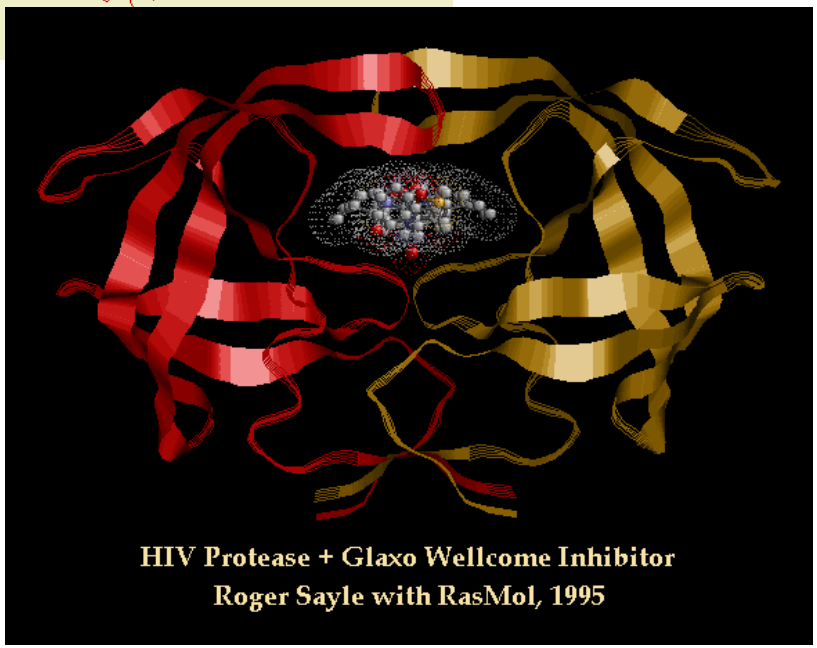


http://alpha2.bmc.uu.se/~kenth/bioinfo/structure/secondary/01.html

Tertiary structure (3D shape)







http://www.umass.edu/microbio/rasmol/sayle1.htm

Folded shape: lowest free energy

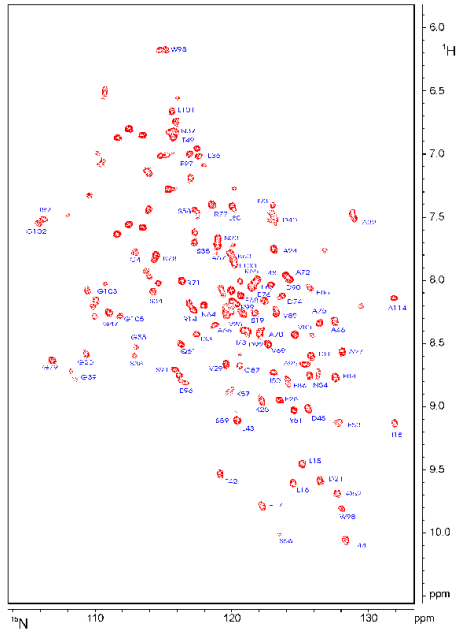
- Energy components
  - electrostatic (~$1/D^2$) (n^2 terms)
  - van der Waals (n^2 terms)
  - hydrogen bonding (n terms)
  - “bending” (n terms)
  - solvent (water/salt) (?? terms)
  - exclusion principle (no two atoms share same volume)
- Energy minimzation
  - small perturbations & computation: hill climbing, simulated annealing, etc.
- Molecular dynamics

How do we know the truth?

- X-ray crystallography
  - crystallize protein
  - shine X-rays
  - examine diffraction patterns

http://www.cryst.bbk.ac.uk/BBS/whatis/cryst_an.html

- Nuclear Magnetic Resonance (NMR)
  - no crystallization necessary
  - magnetic field “vibrates” hydrogen atoms
  - Nobel prize: Kurt Wuethrich



http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm

Simpler problems

- Secondary structure prediction
- Side-chain conformation (assuming fixed backbone)
- Protein docking (how do proteins interact)
- Database searches (protein threading)

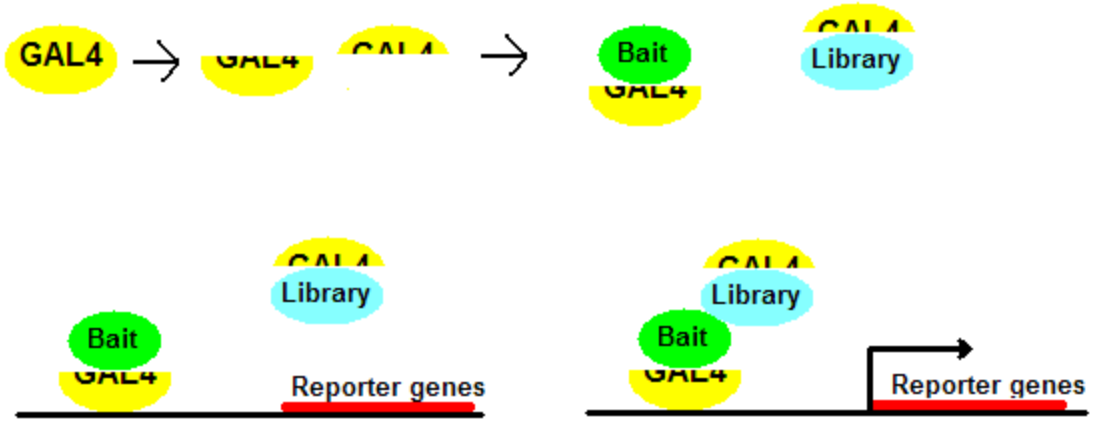
- Simpler energy functions
- Folding on a lattice (theoretical approximation)

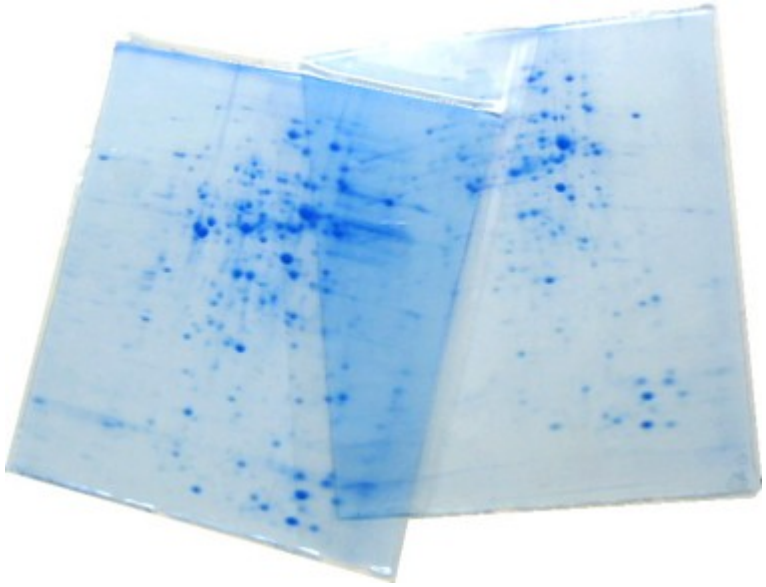
- Critical Assessment of Fully Automated Structure Prediction – competition on proteins with unpublished 3D structure

Proteomics

Proteomics

- Large-scale analysis of proteins
  - protein-protein interactions (e.g. yeast 2-hybrid)
  - 2D gels (mass vs. isoelectric point)
  - Mass-spectrometry
  - Protein microarrays
  - etc.





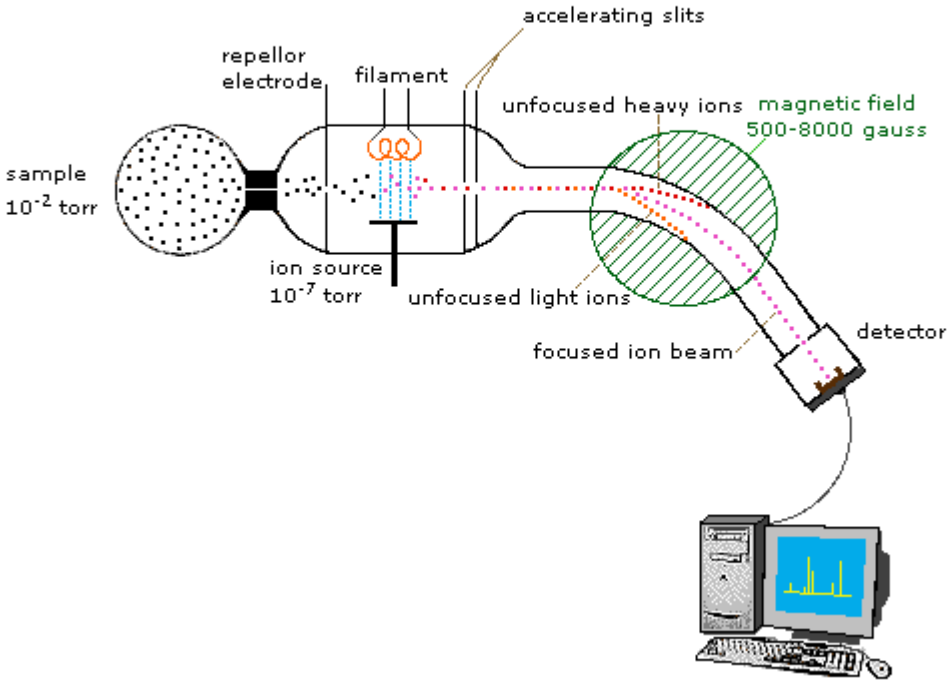
Proteomics

- Why proteomics? Are DNA/RNA microarrays not sufficient?
- RNA abundance is not necessarily related to protein abundance
- Many proteins are modified post-translation
  - addition of additional molecules (phosphate, sugars, etc.)
  - creation of complexes (hemoglobin is actually 4 molecules)

Mass spectrometry

- Technique for measuring the mass-to-charge ratio of ions
- Basic idea
  - shoot ions into a magnetic field
  - deflection depends on mass
- Output of a mass-spectrometer
  - ions “sorted” by mass
  - for each mass bucket - number of ions with that specific mass

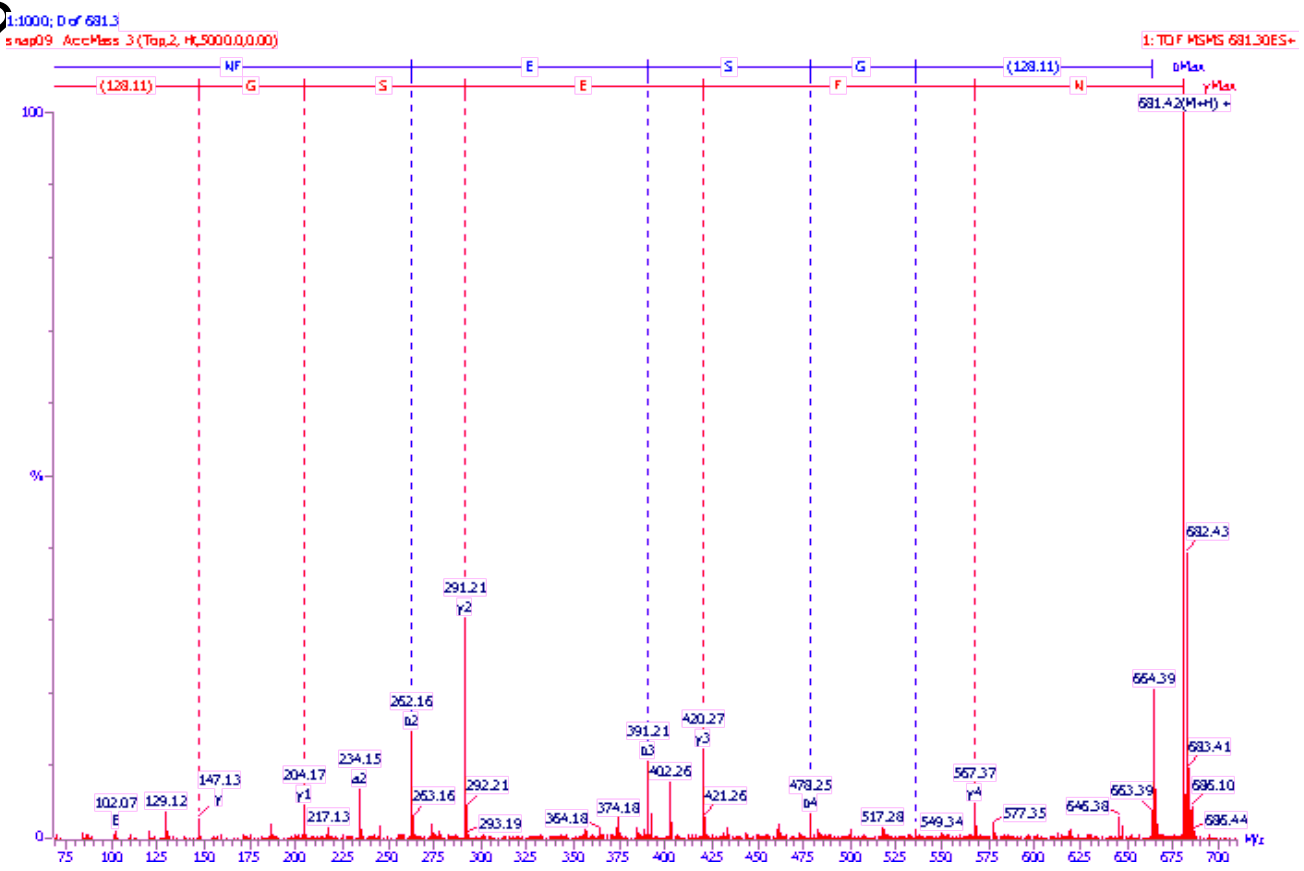
Mass-spectrometry



http://www.cem.msu.edu/~reusch/VirtualText/Spectrpy/MassSpec/masspec1.htm

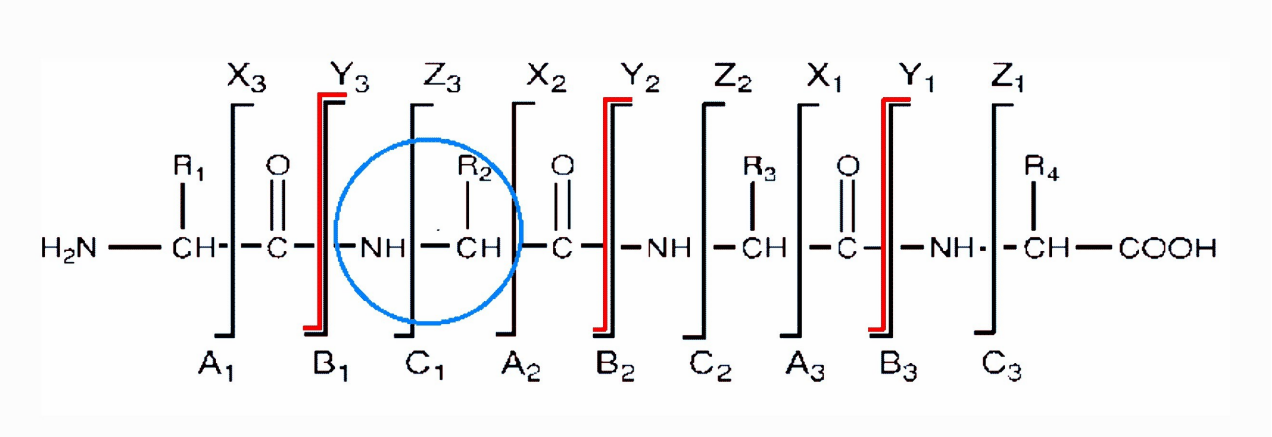
Tandem Mass Spectrometry

- First mass-spectrometer “focuses” on a specific protein
- Second mass-spectrometer breaks the protein into smaller chunks
- Problem: given the chunks, what was the original protein?



Peptide sequencing

- Peptide - a chunk of a protein, usually obtained by enzymatic cleavage of the protein (using trypsin)

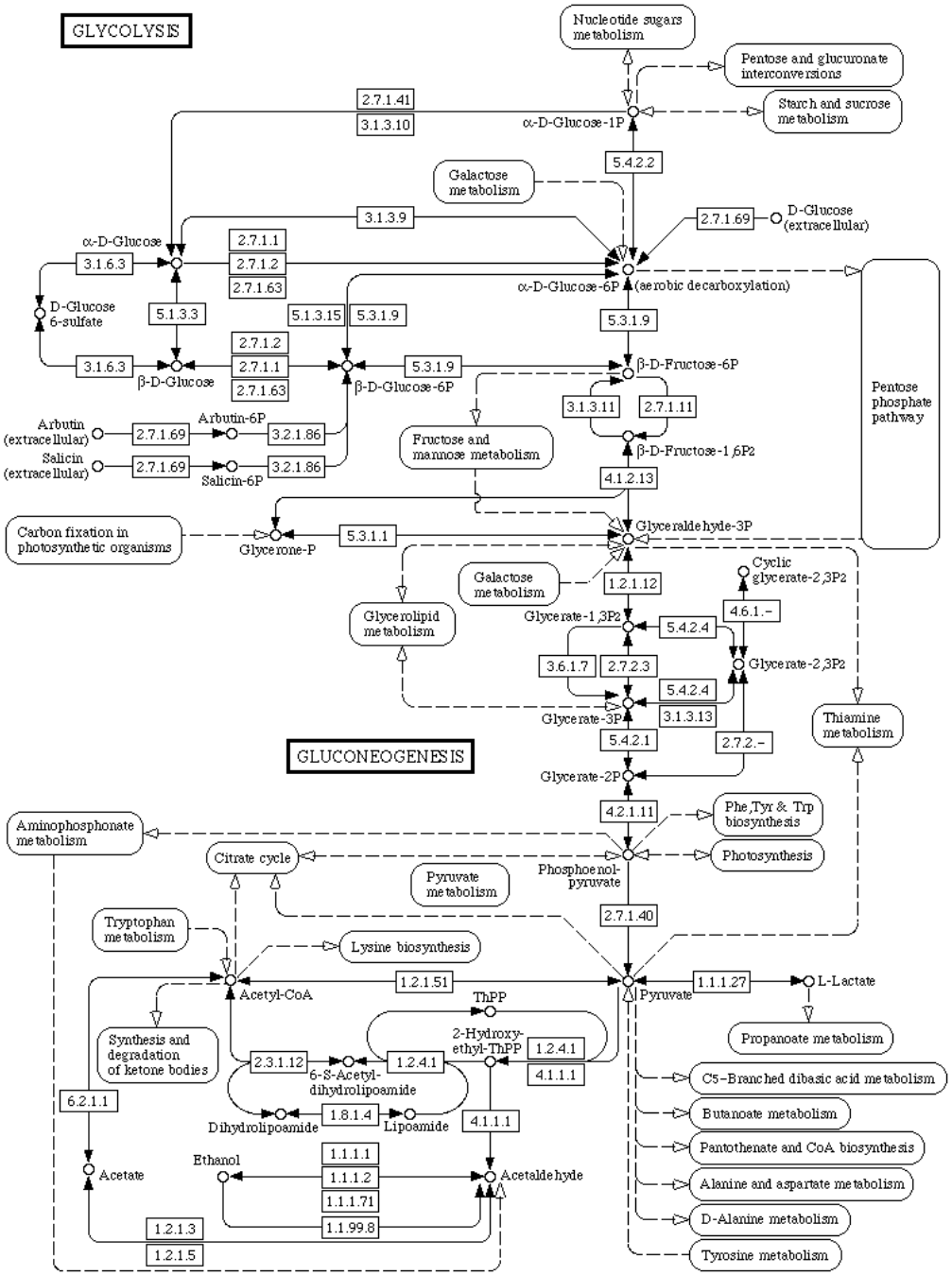


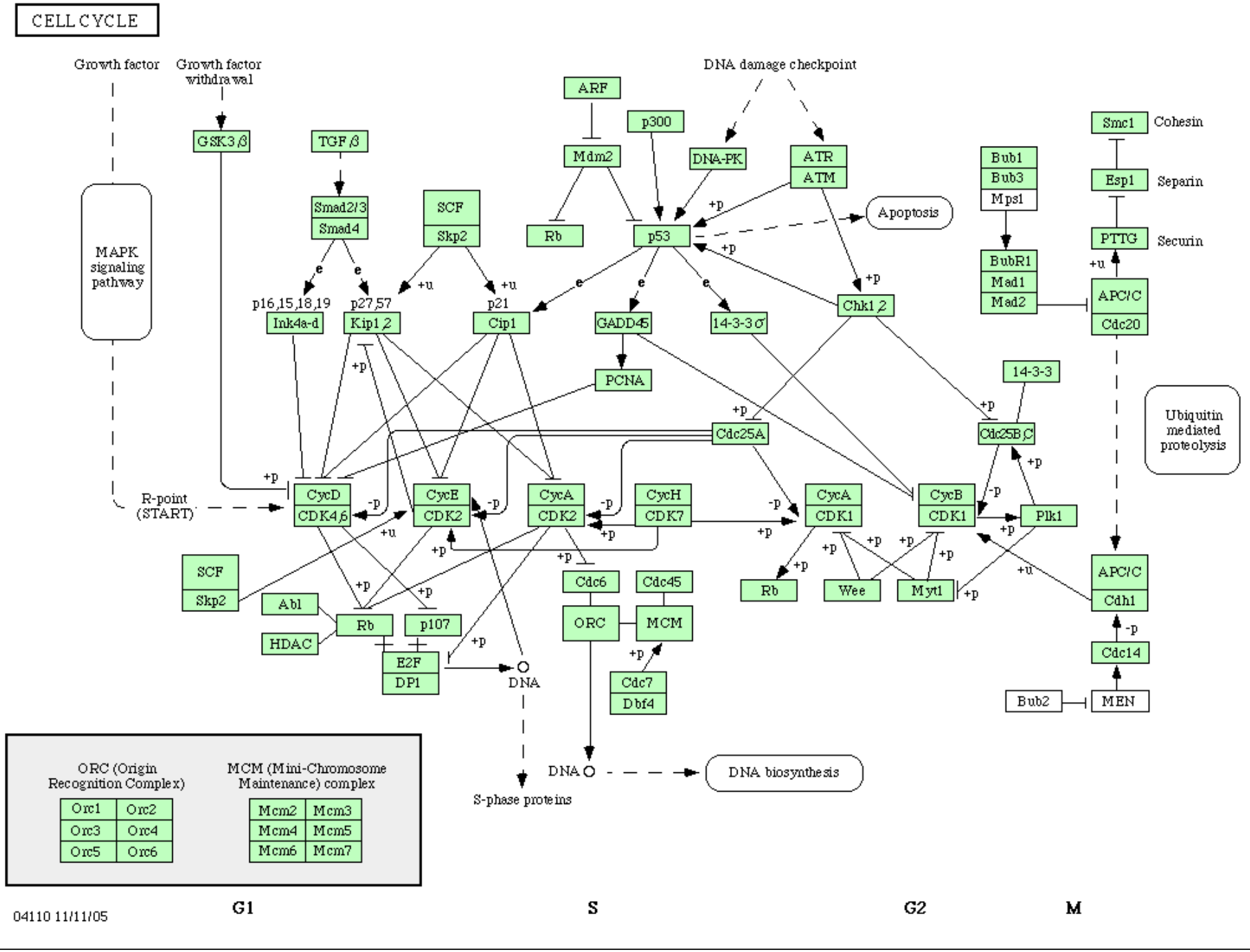
- Problem: Given an MS spectrum (weights of fragments), what was the sequence of the peptide?
- Or: find the peptide (of mass m) that best matches the experimental data

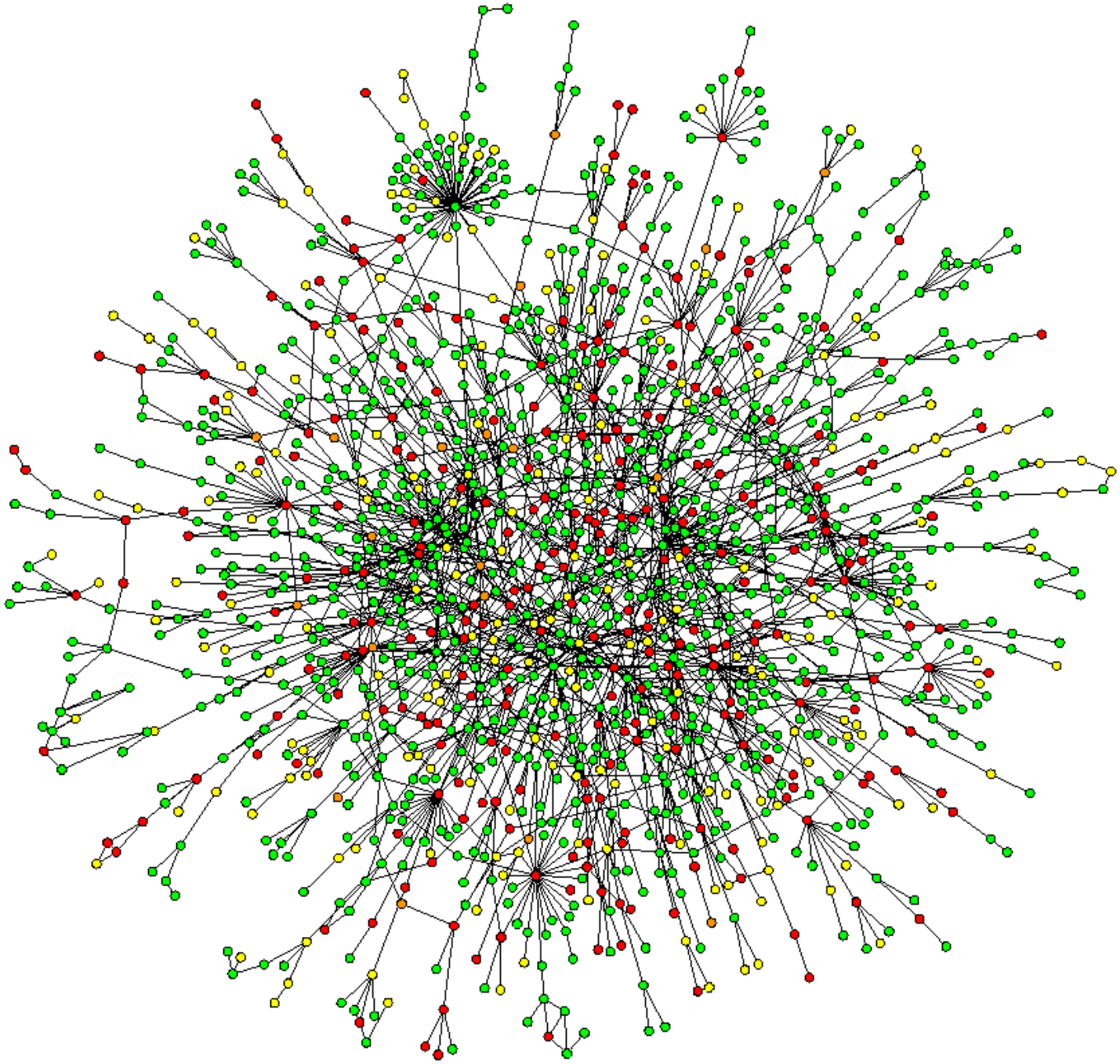
Biological networks

Biological networks

- Genes/proteins do not exist in isolation
- Interactions between genes or proteins can be represented as graphs
- Examples:
  - metabolic pathways
  - regulatory networks
  - protein-protein interactions (e.g. yeast 2-hybrid)
  - genetic interactions (synthetic lethality)


GLYCOLYSIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-1P
5.4.2.2
Galactose metabolism
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
α-D-Glucose
3.1.6.3
2.7.1.1
2.7.1.2
2.7.1.63
α-D-Glucose-6P
(aerobic decarboxylation)
D-Glucose 6-sulfate
5.1.3.3
5.1.3.15
5.3.1.9
5.3.1.9
3.1.6.3
2.7.1.2
2.7.1.1
2.7.1.63
β-D-Glucose
β-D-Glucose-6P
5.3.1.9
β-D-Fructose-6P
Pentose phosphate pathway
3.1.3.11
2.7.1.11
Arbutin (extracellular)
2.7.1.69
Arbutin-6P
3.2.1.86
Salicin (extracellular)
2.7.1.69
Salicin-6P
3.2.1.86
Fructose and mannose metabolism
β-D-Fructose-1,6P2
4.1.2.13
Carbon fixation in photosynthetic organisms
5.3.1.1
Glycerone-P
Glyceraldehyde-3P
Galactose metabolism
1.2.1.12
Cyclic glycerate-2,3P2
4.6.1.-
Glycerolipid metabolism
Glycerate-1,3P2
5.4.2.4
3.6.1.7
2.7.2.3
Glycerate-2,3P2
5.4.2.4
Glycerate-3P
3.1.3.13
Thiamine metabolism
5.4.2.1
2.7.2.-
GLUCONEOGENESIS
Glycerate-2P
4.2.1.11
Phe,Tyr & Trp biosynthesis
Aminophosphonate metabolism
Citrate cycle
Photosynthesis
Pyruvate metabolism
Phosphoenol-pyruvate
2.7.1.40
Tryptophan metabolism
Lysine biosynthesis
1.2.1.51
Acetyl-CoA
Pyruvate
1.1.1.27
L-Lactate
ThPP
2-Hydroxy-ethyl-ThPP
1.2.4.1
1.2.4.1
4.1.1.1
Synthesis and degradation of ketone bodies
2.3.1.12
6-S-Acetyl-dihydrolipoamide
Propanoate metabolism
C5-Branched dibasic acid metabolism
6.2.1.1
Dihydrolipoamide
1.8.1.4
Lipoamide
4.1.1.1
Butanoate metabolism
Pantothenate and CoA biosynthesis
Acetate
Ethanol
1.1.1.1
1.1.1.2
1.1.1.71
1.1.99.8
Acetaldehyde
Alanine and aspartate metabolism
D-Alanine metabolism
1.2.1.3
1.2.1.5
Tyrosine metabolism


CELL CYCLE
Growth factor
Growth factor withdrawal
DNA damage checkpoint
MAPK signaling pathway
GSK3β
TGFβ
Smad2/3
Smad4
SCF
Skp2
ARF
Mdm2
p300
DNA-PK
ATR
ATM
Rb
p53
Apoptosis
Chk1,2
p16,15,18,19
Ink4a-d
p27,57
Kip1,2
p21
Cip1
GADD45
14-3-3σ
PCNA
Cdc25A
R-point
(START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Cdc25B,C
14-3-3
Plk1
Bub1
Bub3
Mps1
BubR1
Mad1
Mad2
Smc1
Cohesin
Esp1
Separin
PTTG
Securin
APC/C
Cdc20
Ubiquitin mediated proteolysis
APC/C
Cdh1
Cdc14
Bub2
MEN
Abl
HDAC
Rb
p107
E2F
DP1
Cdc6
Cdc45
ORC
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
DNA
S-phase proteins
DNA
DNA biosynthesis
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
04110 11/11/05
G1
S
G2
M



Metagenomics

Why do we care?

- Bacteria are everywhere in the environment
- They are not all evil
- Bacteria can be quite useful

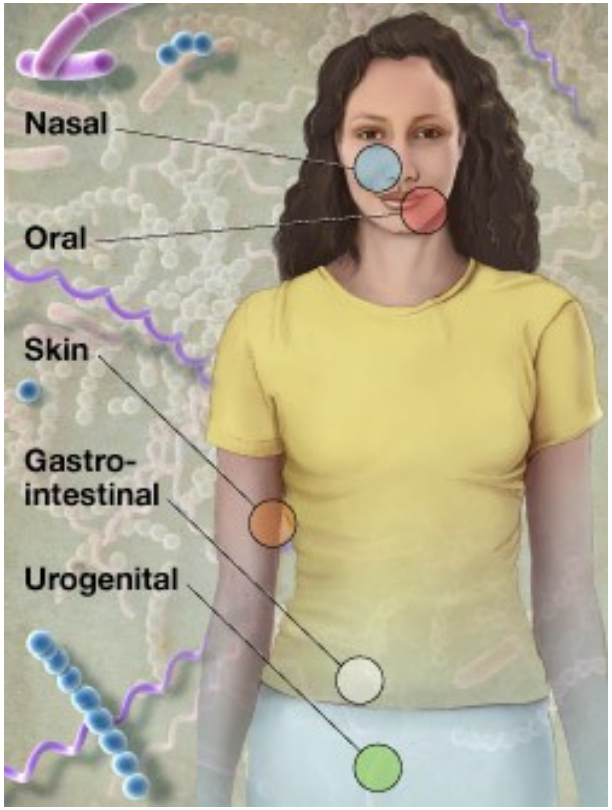
Bio-energy

Bio-remediation

Drug development
antibiotics
anti-cancer

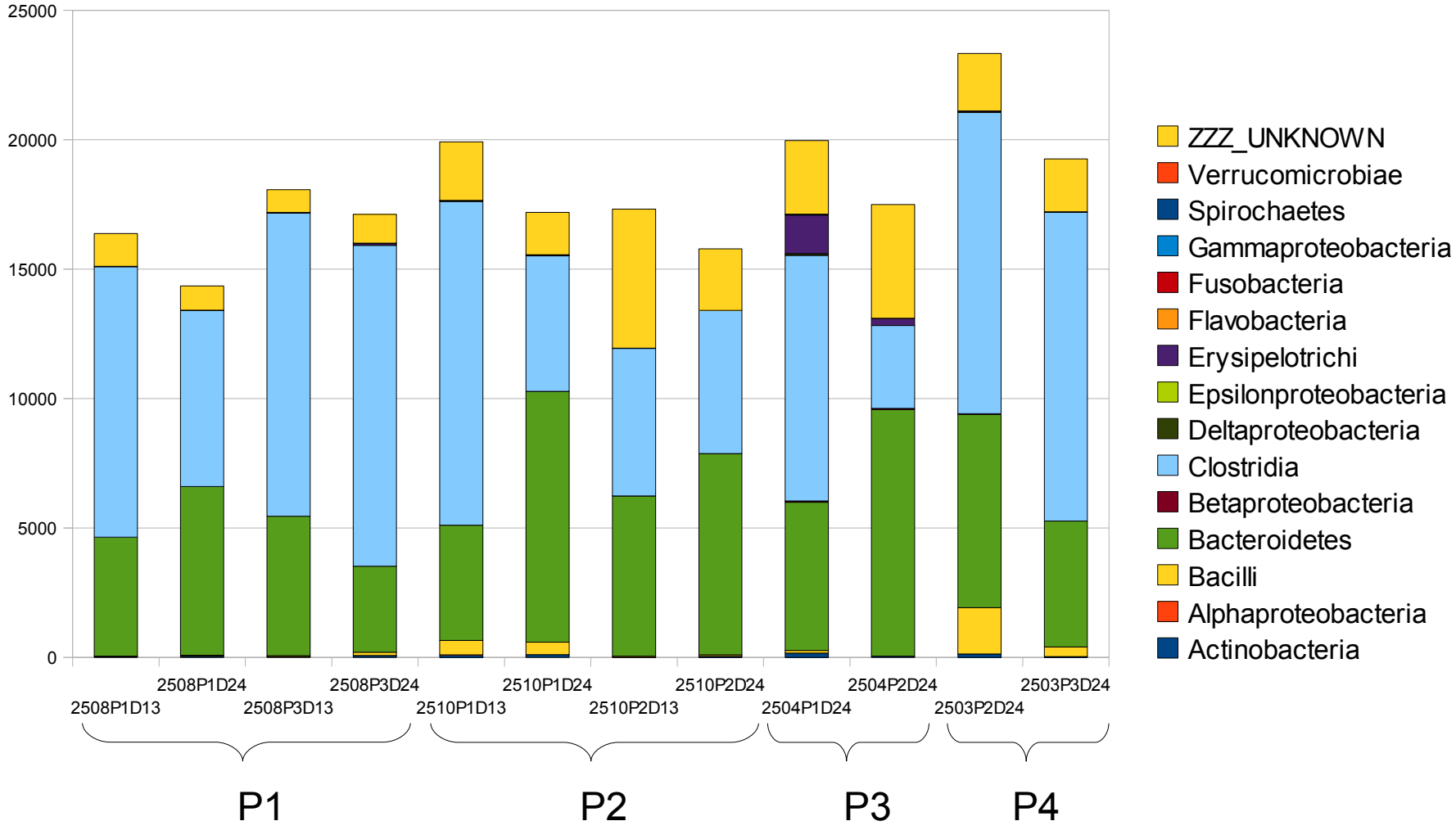
Human microbiome

- Human = 1 order of magnitude more bacterial cells than human cells
  - critical to infant development (immune system, GI-tract)
  - provide essential nutrients (vitamin K, B12, essential amino-acids,....)
  - help digest complex molecules – starches, plant material
  - imbalances in normal bacterial populations correlate with disease (IBD, colon cancer, ...)



Human microbiome project
nihroadmap.nih.gov/hmp/

Some challenges on real data



Spatial genomics

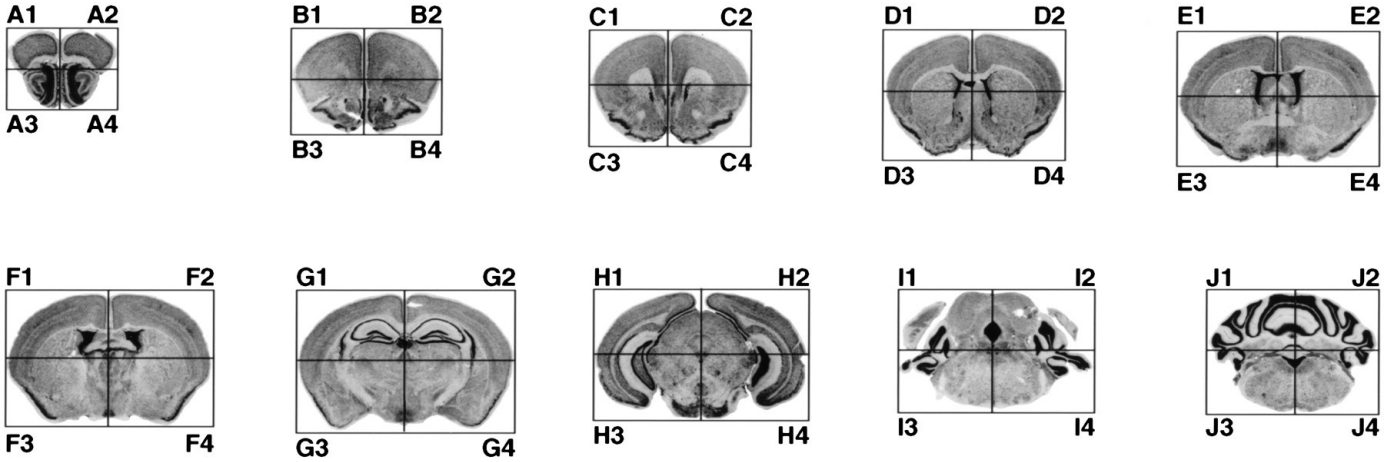
Voxelation

- Brown, V.M., et al., *High-throughput imaging of brain gene expression.* Genome Res, 2002. **12**(2): p. 244-54.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citati

- Brown, V.M., et al., *Multiplex three-dimensional brain gene expression mapping in a mouse model of Parkinson's disease.* Genome Res, 2002. **12**(6): p. 868-84.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citati

- Gene expression information in a spatial context
- Combines microarray analysis with computer graphics

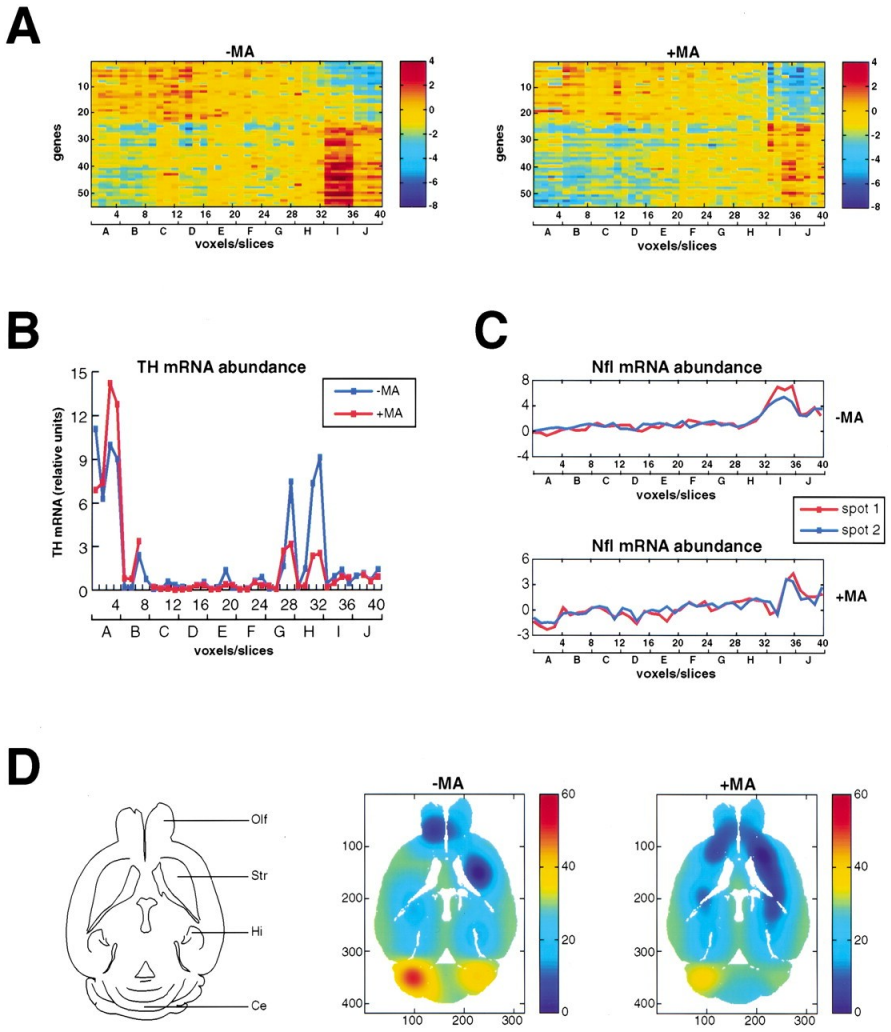
Figure 2 Voxelation scheme



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

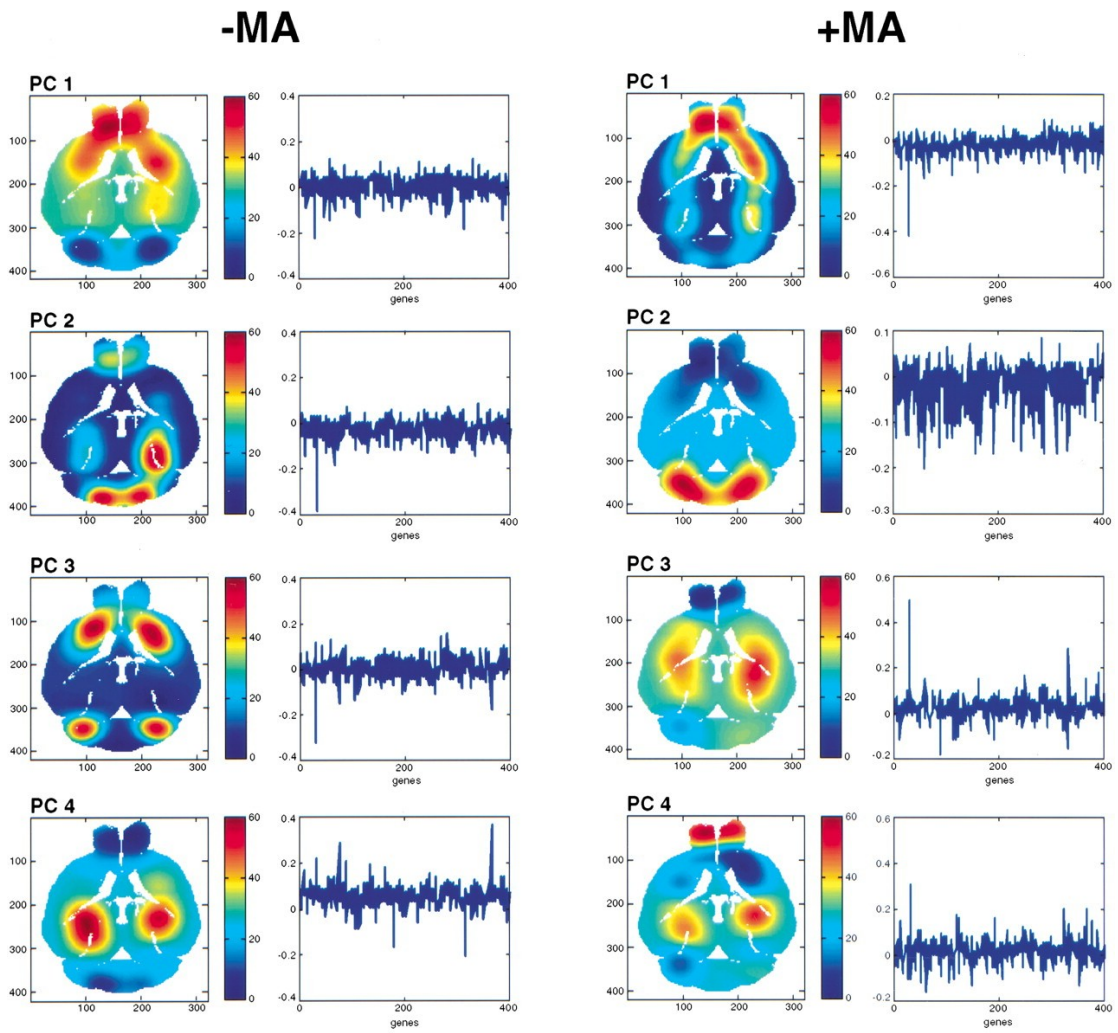
- Mouse brain cut up into voxels
- Run a separate microarray experiment on each voxel

Figure 4 Spatial gene expression patterns for the subset of correlated genes



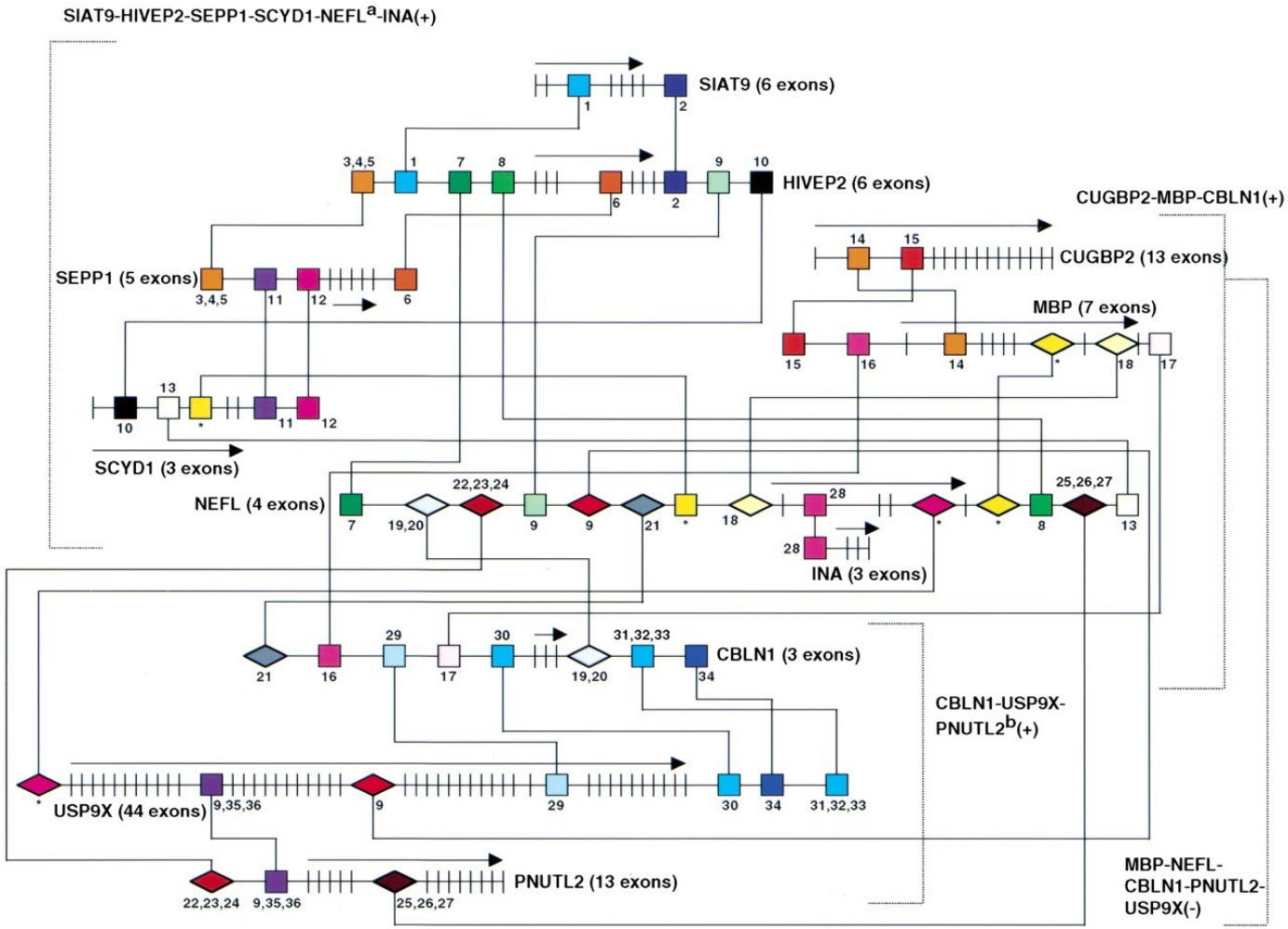
Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Figure 7 SVD delineates anatomical regions of the brain



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Figure 5 Putative regulatory elements shared between groups of correlated and anticorrelated genes



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884