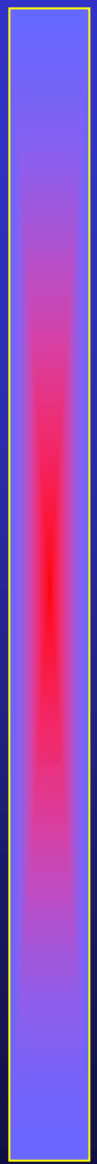
DA SEQUENCIA GENOMICA …. À GENOMICA FUNCIONAL E APLICADA

Slides por: Wim Degrave
Leila de Mendonça Lima
Antonio B. de Miranda

Departamente de Bioquímica e
Biologia Molecular
Instituto Oswaldo Cruz - Fiocruz
Rio de Janeiro, Brasil
wdegrave@fiocruz.br
http://www.dbbm.fiocruz.br

BIOLOGIA



Análise de sequências nucleotídicas e proteicas
Montagem de sequenciamento shotgun em “contigs” e genomas
Anotação de genomas, clusterização e datamining
Genoma funcional
Proteoma
Microarrays
Comparação de genomas
Modelagem molecular, desenvolvimento de drogas, docking
Evolução molecular e filogenia
Bioestatística, genética de populações
Estruturação de bancos de dados, métodos para datamining
Interfaces, scripting e visualização
Desenvolvimento e otimização de algoritmos
Processamento paralelo, Grids

INFORMÁTICA

Annotation is the description of:

- Function(s) of the protein
- Post-translational modification(s)
- Domains and sites
- Secondary structure
- Quaternary structure
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

Additional information for proteins

- ALTERNATIVE PRODUCTS
- CATALYTIC ACTIVITY
- COFACTOR
- DEVELOPMENTAL STAGE
- DISEASE
- DOMAIN
- ENZYME REGULATION
- FUNCTION
- INDUCTION
- PATHWAY
- PHARMACEUTICALS
- POLYMORPHISM
- PTM
- SIMILARITY
- SUBCELLULAR LOCATION
- SUBUNIT
- TISSUE SPECIFICITY

Amino-acid sites are:

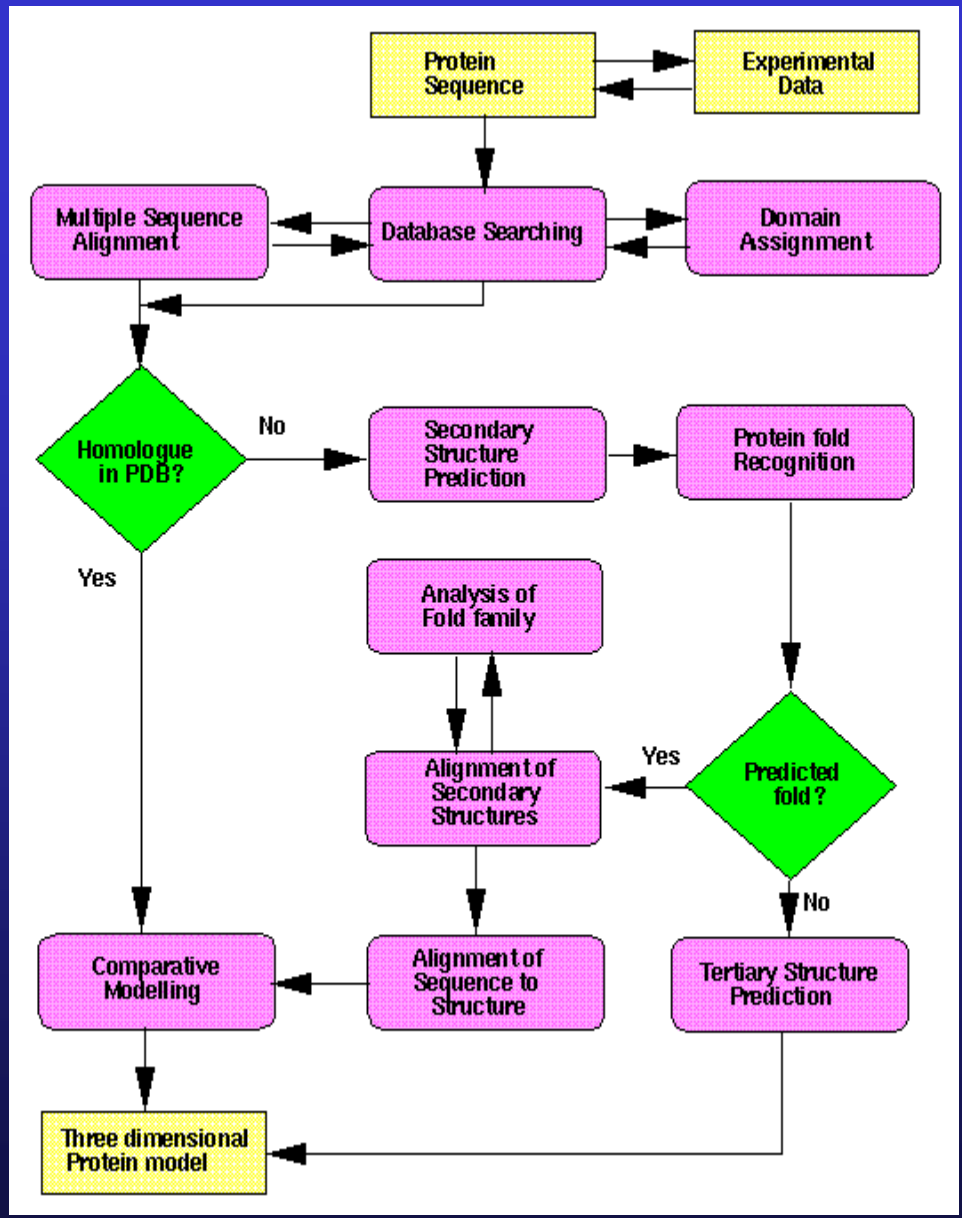
- Post-translational modification of a residue
- Covalent binding of a lipidic moiety
- Disulfide bond
- Thiolester bond
- Thioether bond
- Glycosylation site
- Binding site for a metal ion
- Binding site for any chemical group (co-enzyme, prosthetic group, etc.)

Regions:

- SIGNAL SEQUENCE
- TRANSIT PEPTIDE
- PROPEPTIDE
- CHAIN
- PEPTIDE
- DOMAIN
- ACTIVE SITE
- DNA BIND SITE
- METAL BIND SITE
- MOLECULE BIND SITE
- TRANSMEMBRANE

Approaches to functional annotation:

- Automatic annotation (sequence homology, rules, transfer info from pdb)
- Automatic classification (pattern databases, clustering, structure)
- Automatic characterisation (functional databases)
- Context information (comparitive genome analysis, metabolic pathway databases)
- Experimental results (2D gels, microarrays)
- Full manual annotation (SWISS-PROT style)



From sequence to function

Predicting function from sequence similarity

- **Orthologues**- arose from speciation, same gene in different organisms -can have <30% homology
- **Paralogues**- from duplication within a genome, second copy may have new or changed function

(difficult to distinguish between otho- and paralogues unless whole genome is available)

- **Equivalog**- proteins with equivalent functions
- **Analog**- proteins catalyzing same reaction but not structurally related
- Some enzymes may have seq similarity simply because common catalytic site, substrate, pathway.

Example of a partial entry

```
ID   Q24469       PRELIMINARY;      PRT;   611 AA.
AC   Q24469; Q24470; Q9VEY7;
DT   01-NOV-1996 (TrEMBLrel. 01, Created)
DT   01-JUN-2000 (TrEMBLrel. 14, Last sequence update)
DT   01-MAR-2001 (TrEMBLrel. 16, Last annotation update)
DE   RAC protein kinase DRAC-PK85.
GN   Akt1{EI6,EP8} OR CG4006{EP8}.
OS   Drosophila melanogaster (Fruit fly){EP7}.
..
OX   NCBI_TaxID=7227;
RN   [1]{EC3}
RP   SEQUENCE FROM N.A. (SHORT ISOFORM).
RC   STRAIN=BERKELEY;
RX   MEDLINE=20196006; PubMed=10731132;
RA   Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D.,
..
RT   "The genome sequence of Drosophila melanogaster.";
RL   Science 287:2185-2195(2000).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=95181376; PubMed=7876156;
RA   Andjelkovic M., Jones P.F., Grossniklaus U., Cron P., Schier A.
RA   Dick M., Bilbe G., Hemmings B.A.;
RT   "Developmental regulation of expression and activity of
RT   multiple forms of the Drosophila RAC protein kinase.";
RL   J. Biol. Chem. 270:4066-4075(1995).
CC   -!- ALTERNATIVE PRODUCTS: 2 ISOFORMS; A LONG FORM (SHOWN HERE)
CC       SHORT FORM; ARE PRODUCED BY ALTERNATIVE SPLICING{EC3}.
CC   -!- SIMILARITY: TO THE SER/THR FAMILY OF PROTEIN KINASES{EA2}.
DR   EMBL; AE003711; AAF55276.1; -.{EI4}
DR   EMBL; X83510; CAA58499.2; -.
DR   EMBL; X83510; CAA58500.1; -.
DR   HSSP; Q63450; 1A06.{EI5}
DR   FlyBase; FBgn0010379; Akt1.{EI6}
DR   InterPro; IPR000719; -.
DR   InterPro; IPR000961; -.
DR   InterPro; IPR001849; -.
DR   InterPro; IPR002290; -.
DR   Pfam; PF00069; pkinase; 1.
DR   Pfam; PF00169; PH; 1.
DR   Pfam; PF00433; pkinase_C; 1.
DR   PROSITE; PS50003; PH_DOMAIN; 2.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; 1.
DR   PROSITE; PS50011; PROTEIN_KINASE_DOM; 1.
DR   PROSITE; PS00108; PROTEIN_KINASE_ST; 1.
DR   SMART; SM00233; PH; 1.
DR   SMART; SM00133; S_TK_X; 1.
KW   ATP-binding{EA1}; Alternative splicing{EC3};
KW   Serine/threonine-protein kinase{EA2}; Transferase{EA1}.
FT   VARSPLIC      1     81       MISSING (IN ISOFORM SHORT){EC3}.
**
**   ##################    INTERNAL SECTION    ##################
**EV EA1; Rulebase; -; RU000304; 22-JAN-2001.
**EV EA2; Rulebase; -; RU000305; 22-JAN-2001.
**EV EC3; Curator; ELW; -; 11-JAN-2001.
**EV EI4; EMBL; -; AAF55276.1; 17-OCT-2000.
**EV EI5; HSSP_ADD; -; Q63450; 29-SEP-2000.
**EV EI6; FLYBASE_ADD; -; FBgn0010379; 28-NOV-2000.
**EV EP7; TREMBL; -; AAF55276.1; 17-OCT-2000.
**EV EP8; DROMEfix; -; v1.2; 18-DEC-2000.
SQ   SEQUENCE   611 AA;  68514 MW;  C1393E43CC27AC34 CRC64;
..
//
```

Life Defined

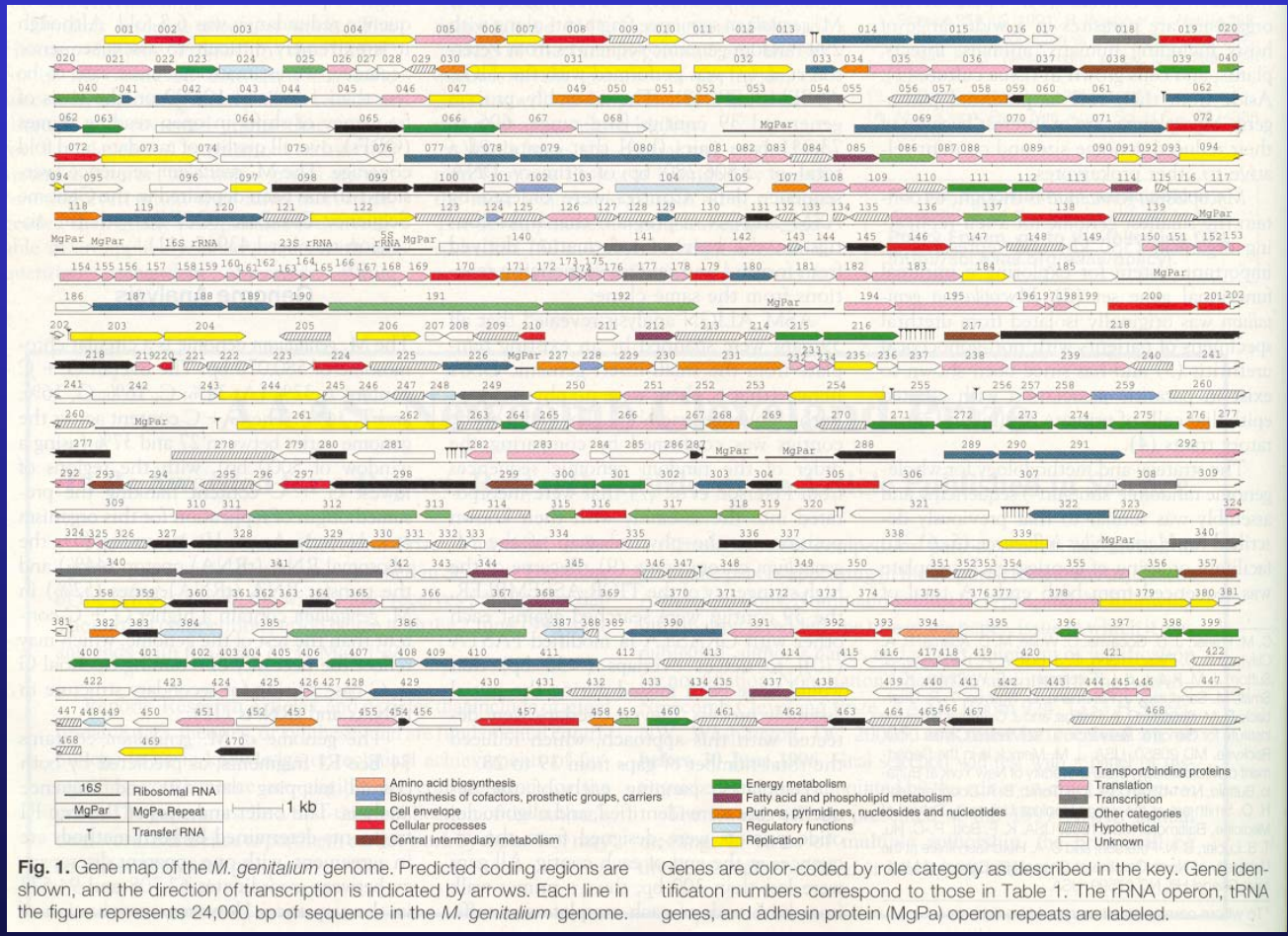


Fig. 1. Gene map of the *M. genitalium* genome. Predicted coding regions are shown, and the direction of transcription is indicated by arrows. Each line in the figure represents 24,000 bp of sequence in the *M. genitalium* genome. Genes are color-coded by role category as described in the key. Gene identificaton numbers correspond to those in Table 1. The rRNA operon, tRNA genes, and adhesin protein (MgPa) operon repeats are labeled.

The complete genome of *M. genitalium*
Fraiser *et al*, Science 270, 397 (1995)

METABOLISM (177 entries)
- amino-acid metabolism (37 entries)
  - amino-acid biosynthesis (33 entries)
  - regulation of amino-acid metabolism (11 entries)
  - amino-acid transport (0 entries)
  - amino-acid degradation (catabolism) (10 entries)
  - other amino-acid metabolism activities (0 entries)
- nitrogen and sulphur metabolism (27 entries)
  - nitrogen and sulphur utilization (21 entries)
  - regulation of nitrogen and sulphur utilization (9 ent
  - nitrogen and sulphur transport (0 entries)
- nucleotide metabolism (42 entries)
  - purine-ribonucleotide metabolism (10 entries)
  - pyrimidine-ribonucleotide metabolism (14 entries)
  - deoxyribonucleotide metabolism (6 entries)
  - metabolism of cyclic and unusual nucleotides (9 en

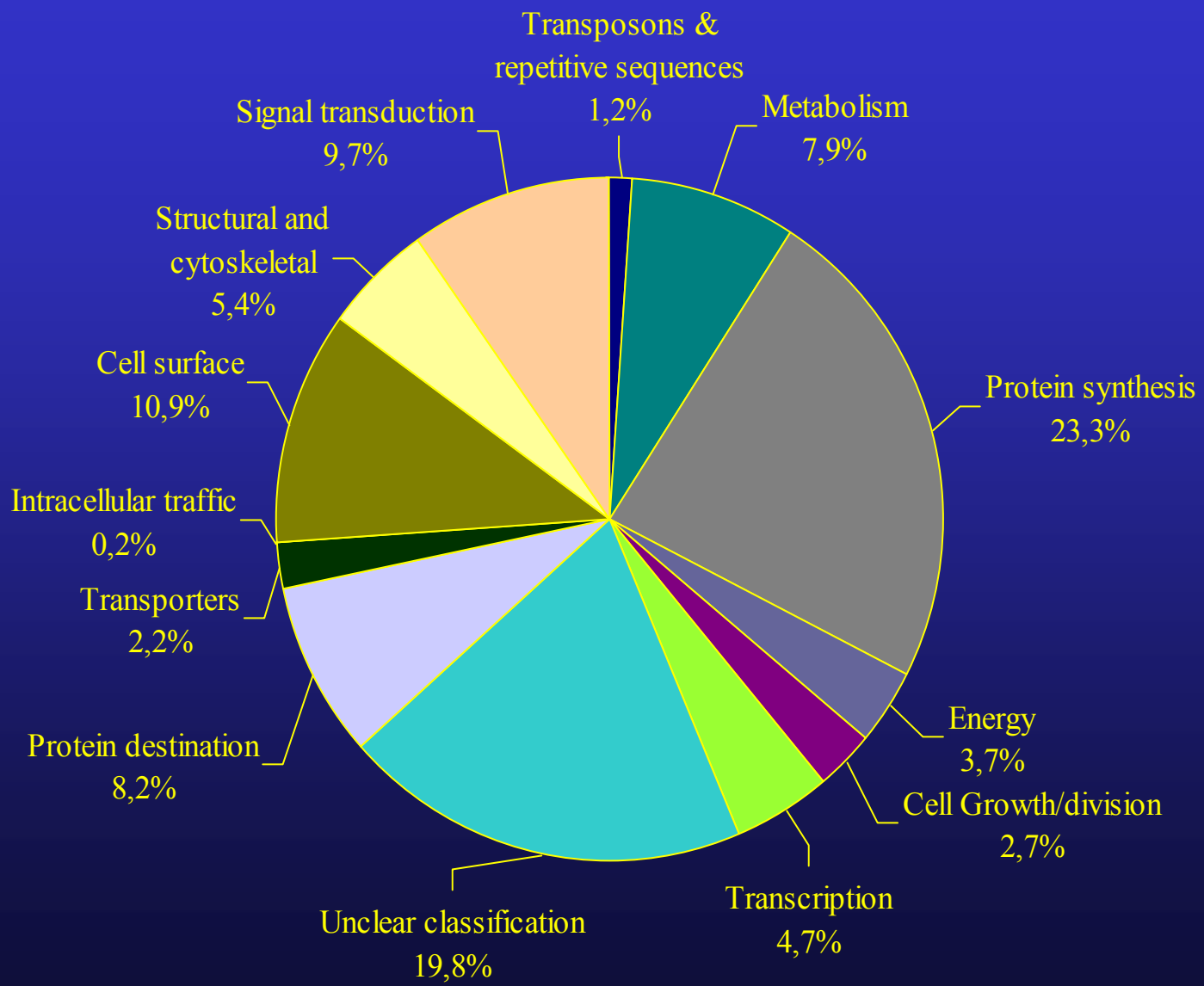
1. -. -.- Oxidoreductases.
1. 1. -.- Acting on the CH-OH group of donors.
1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
1. 1. 2.- With a cytochrome as acceptor.
1. 1. 3.- With oxygen as acceptor.
1. 1. 4.- With a disulfide as acceptor.
1. 1. 5.- With a quinone or similar compound as acceptor.
1. 1.99.- With other acceptors.
1. 2. -.- Acting on the aldehyde or oxo group of donors.
1. 2. 1.- With NAD(+) or NADP(+) as acceptor.
1. 2. 2.- With a cytochrome as acceptor.
1. 2. 3.- With oxygen as acceptor.
1. 2. 4.- With a disulfide as acceptor.
1. 2. 7.- With an iron-sulfur protein as acceptor.
1. 2.99.- With other acceptors.
1. 3. -.- Acting on the CH-CH group of donors.

- ATPases (13 entries)
- Glycoproteins (31 entries)
- Heat shock proteins (24 entries)
- Histones (31 entries)
- Hydrogenases (26 entries)
- Isomerases (21 entries)
- Kinases (17 entries)
- Kinetoplast (3 entries)
- Phosphatases (5 entries)
- Polymerases (1 entries)
- Proteases/proteinases (58 entries)
- Ribosomal (37 entries)

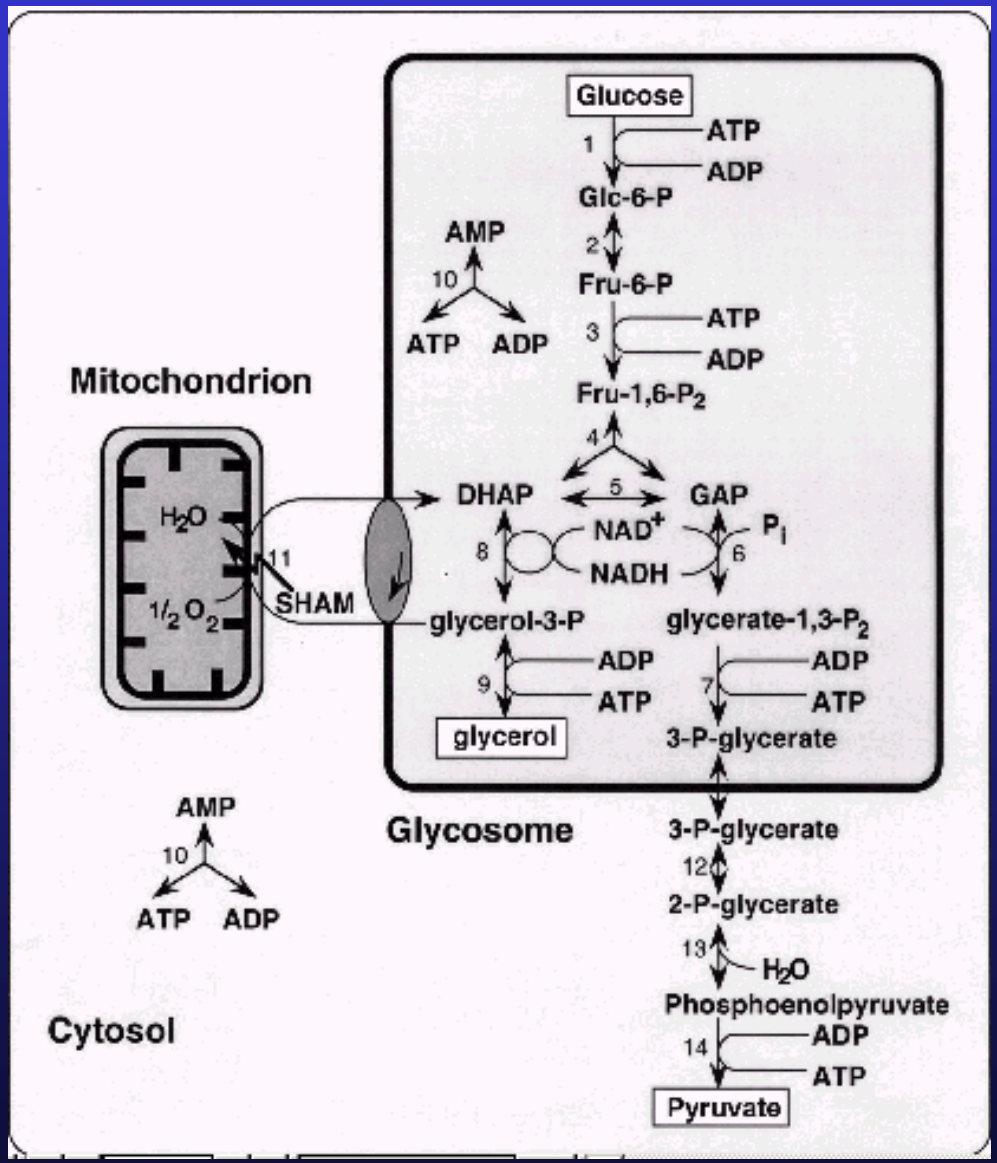
- Glycolysis Gluconeogenesis (17 entries)
  - Citrate cycle TCA cycle (12 entries)
  - Pentose phosphate cycle (1 entries)
  - Pentose and glucuronate interconversions (0 entries)
  - Fructose and mannose metabolism (2 entries)
  - Galactose metabolism (0 entries)
  - Ascorbate and aldarate metabolism (0 entries)
  - Fatty acid biosynthesis path 1 (0 entries)
  - Fatty acid biosynthesis path 2 (1 entries)
  - Fatty acid metabolism (1 entries)

Functional classification of T. cruzi ESTs

(Verdun *et al.*, 1998)



What genes and pathways were uncovered thusfar?



Opperdoes & Michels, Int. J. Parasitol. 31, 482-490, 2001

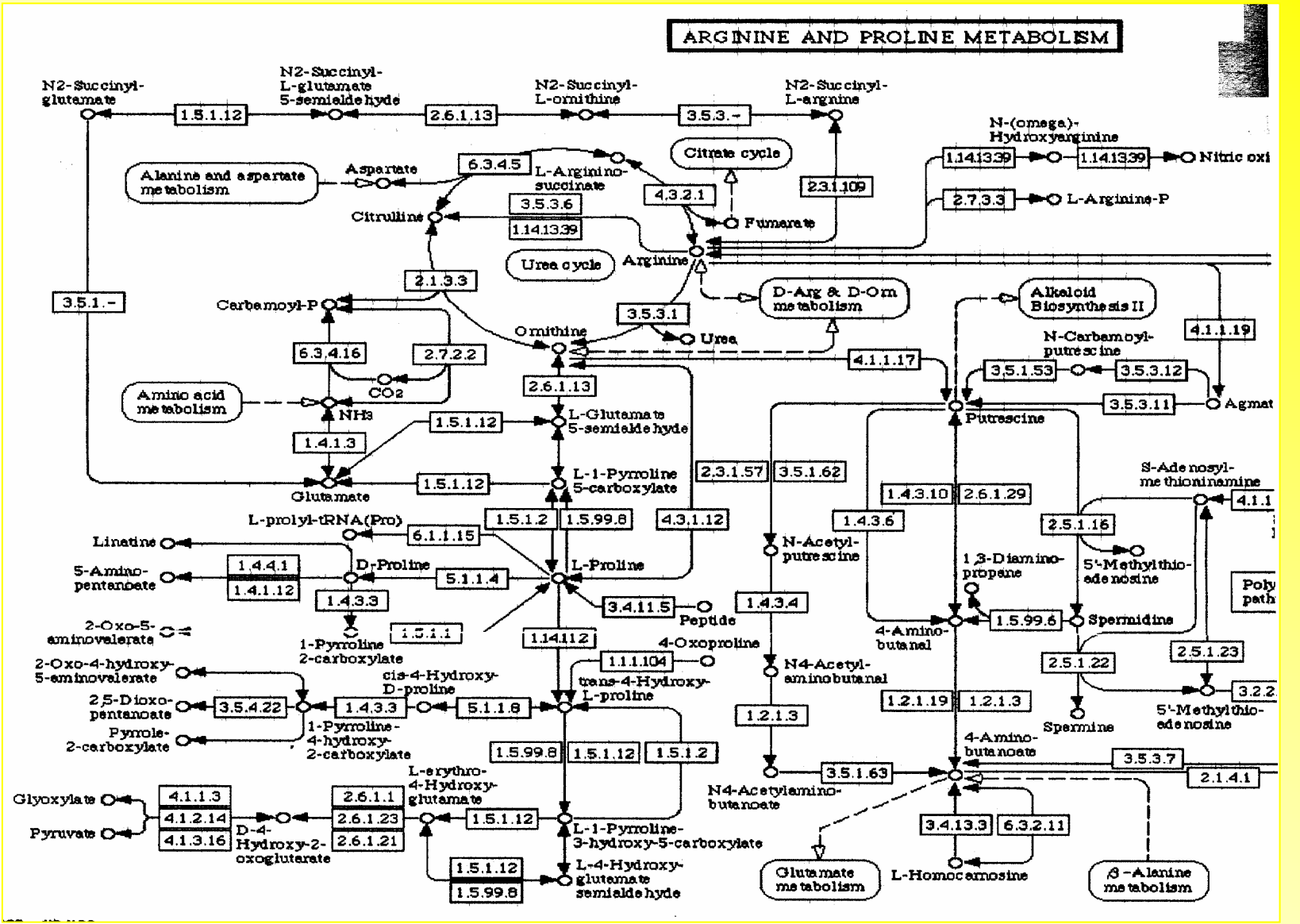
Ex. Glycolysis :

1. Hexokinase (4 EST/GSS)
2. Phosphoglucose isomerase
3. Phosphofructokinase (1)
4. Aldolase (4 EST/GSS)
5. Triose phosphate isomerase (2EST+ 1 GB)
6. Glyceraldehyde 3 phosphate dehydrogenase (5 EST/GSS + 1 GB)
7. Phosphoglycerate kinase
8. Glycerol 3 phosphate dehydrogenase
9. Glycerol kinase
10. Adenylate kinase
11. Glycerol 3 phosphate oxidase
12. Phosphoglycerate mutase (1 EST)
13. Enolase (3) (3 EST/GSS + 3 GB)
14. Pyruvate kinase

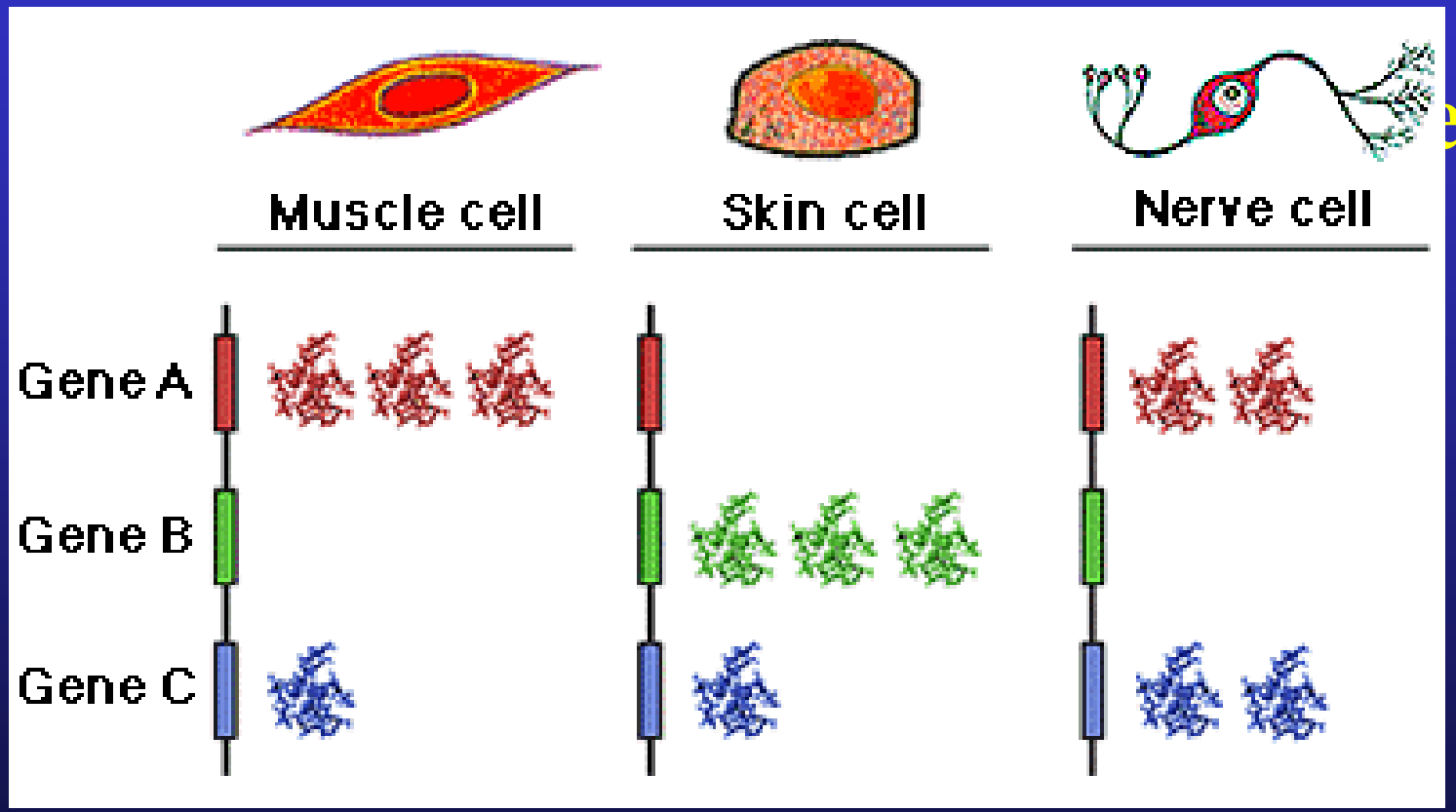
annotated Found upon search

pentose pathway : 7 EST/GSS, 4/8 enz.

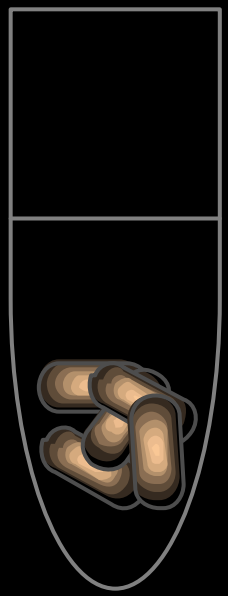
General Proline and Arginine Metabolism



Tissue specific gene expression

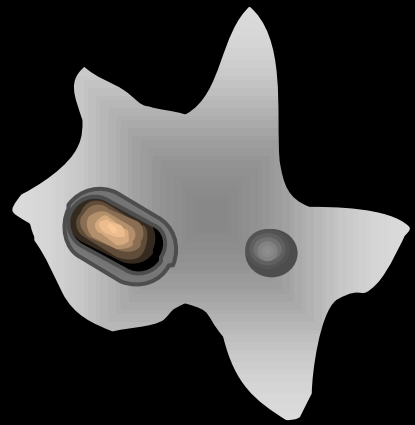


Perfil comparativo de expressão gênica



Meio axênico

X



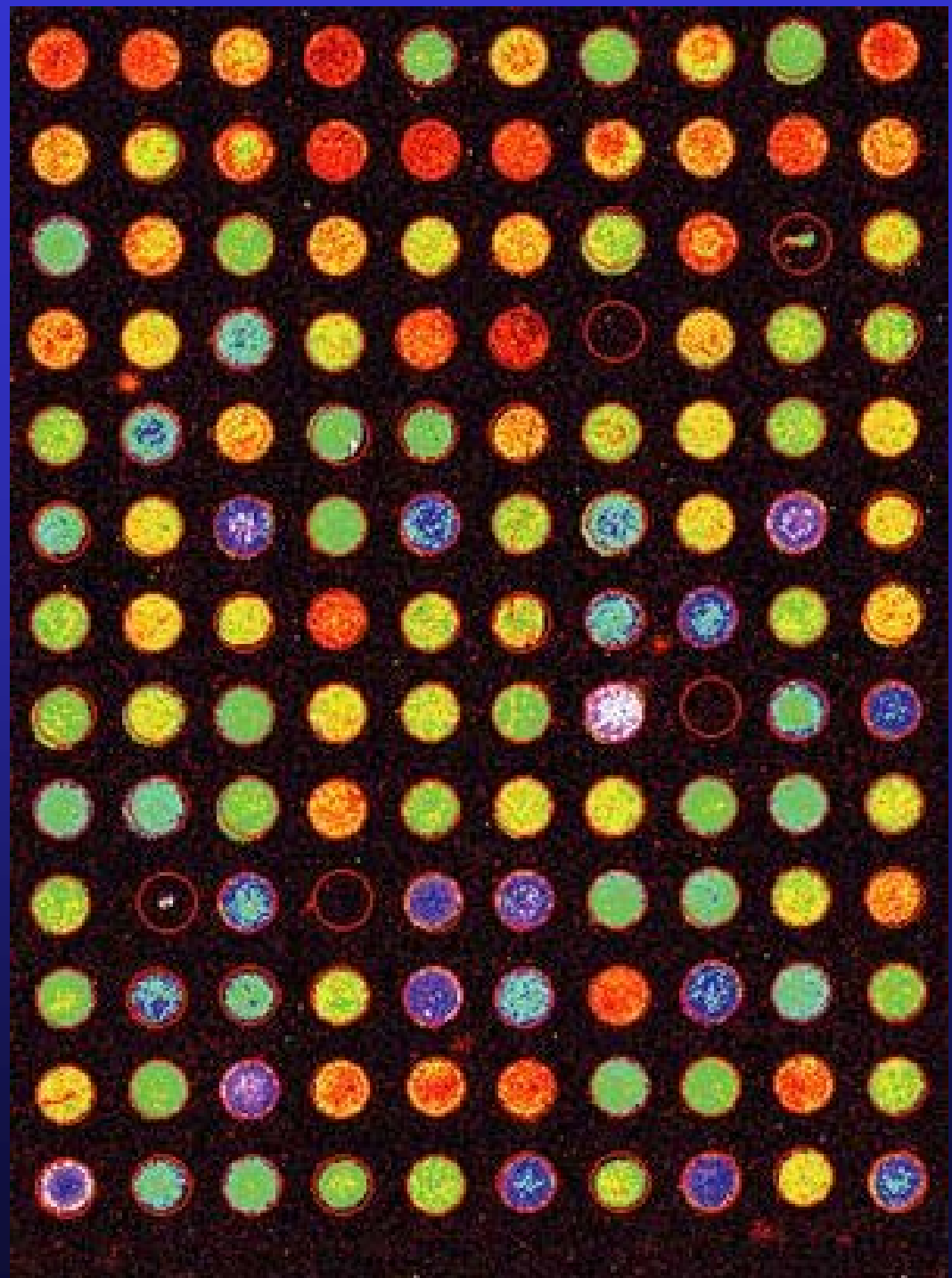
in vitro

X

in vivo

Patógeno + hospedeiro

DNA microarrays:

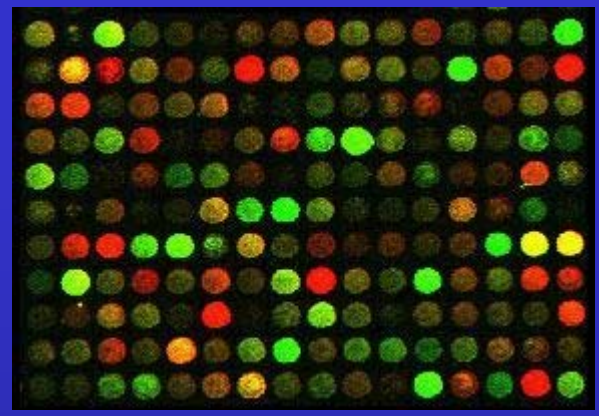


DNA Chips

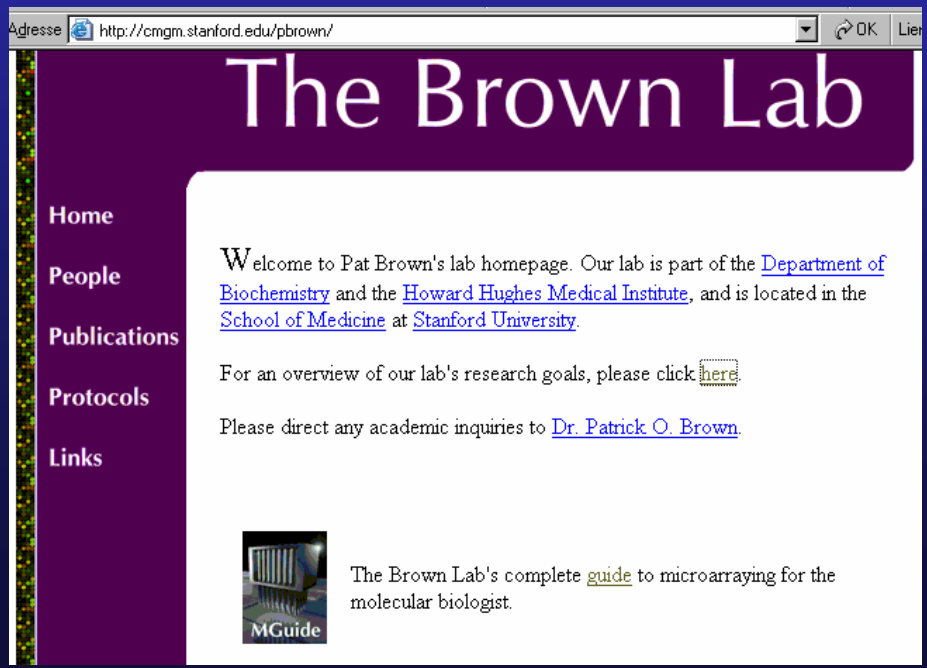
STEPHEN P. A. FODOR (1953-)

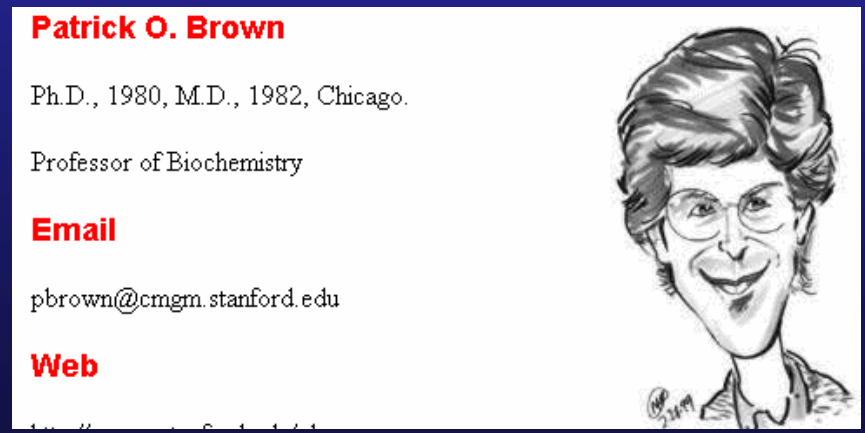
Fodor fez pós-doutorado na UC Berkeley e em 1989 foi trabalhar para a Affymax Research Institute em Palo Alto. Ele era responsável por desenvolver um processo para gerar arranjos (*arrays*) de alta densidade, miniaturizados, contendo compostos biológicos. Isso levou ao desenvolvimento do primeiro *chip* de DNA e das técnicas necessárias para ler e analisar os resultados para estudos genômicos em larga escala. O processo foi bastante refinado desde então e o sucesso desta tecnologia levou, em 1993, à criação de outra empresa, a Affymatrix Inc. Fodor é atualmente o CEO da Affymatrix.

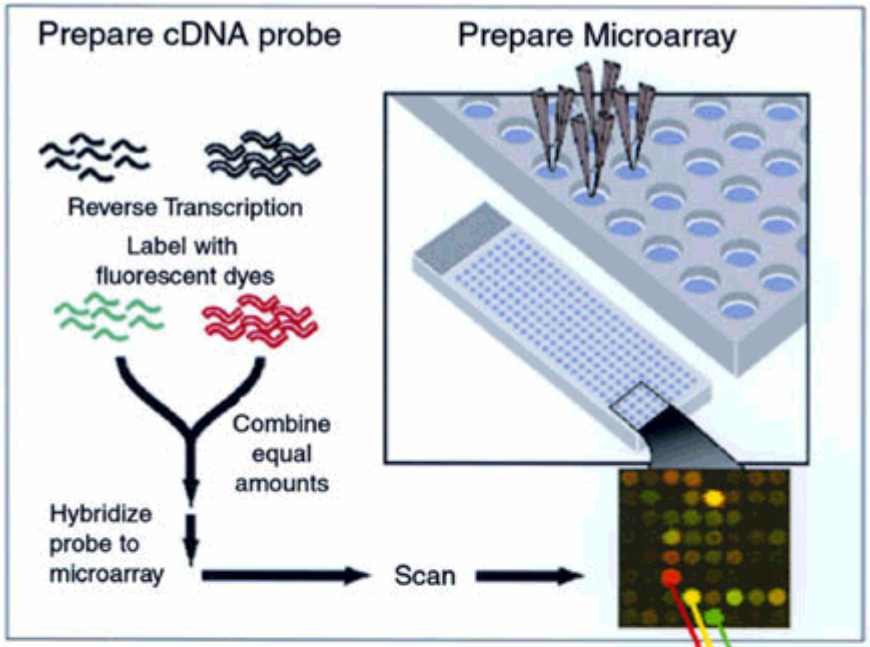
1995 Stanford University - Pat Brown and Ron Davis



Desenvolveram um método para preparar chips de DNA no laboratório, "spotando" fragmentos inteiros de DNA em uma lâmina de microscópio. Também construíram um robô que utiliza tips para aplicar as gotinhas contendo DNA. O sítio web do laboratório em Stanford fornece instruções completas para a confecção de chips de DNA, tornando essa tecnologia acessível à comunidade científica.

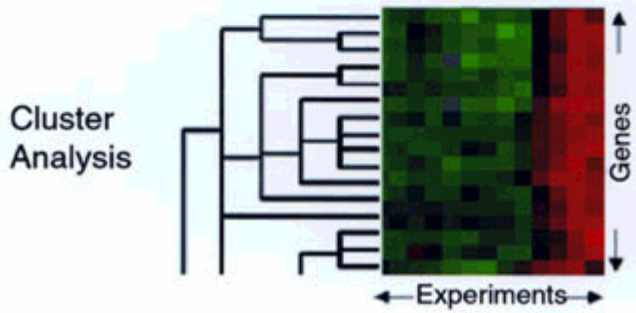






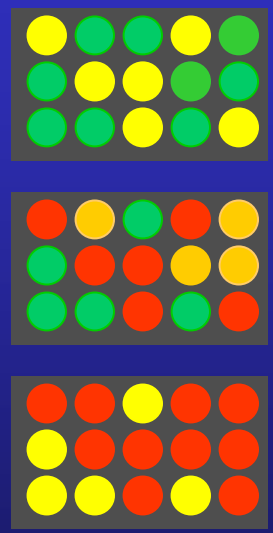
Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2 \frac{Cy5}{Cy3}$
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91

Image Analysis



Cluster Analysis

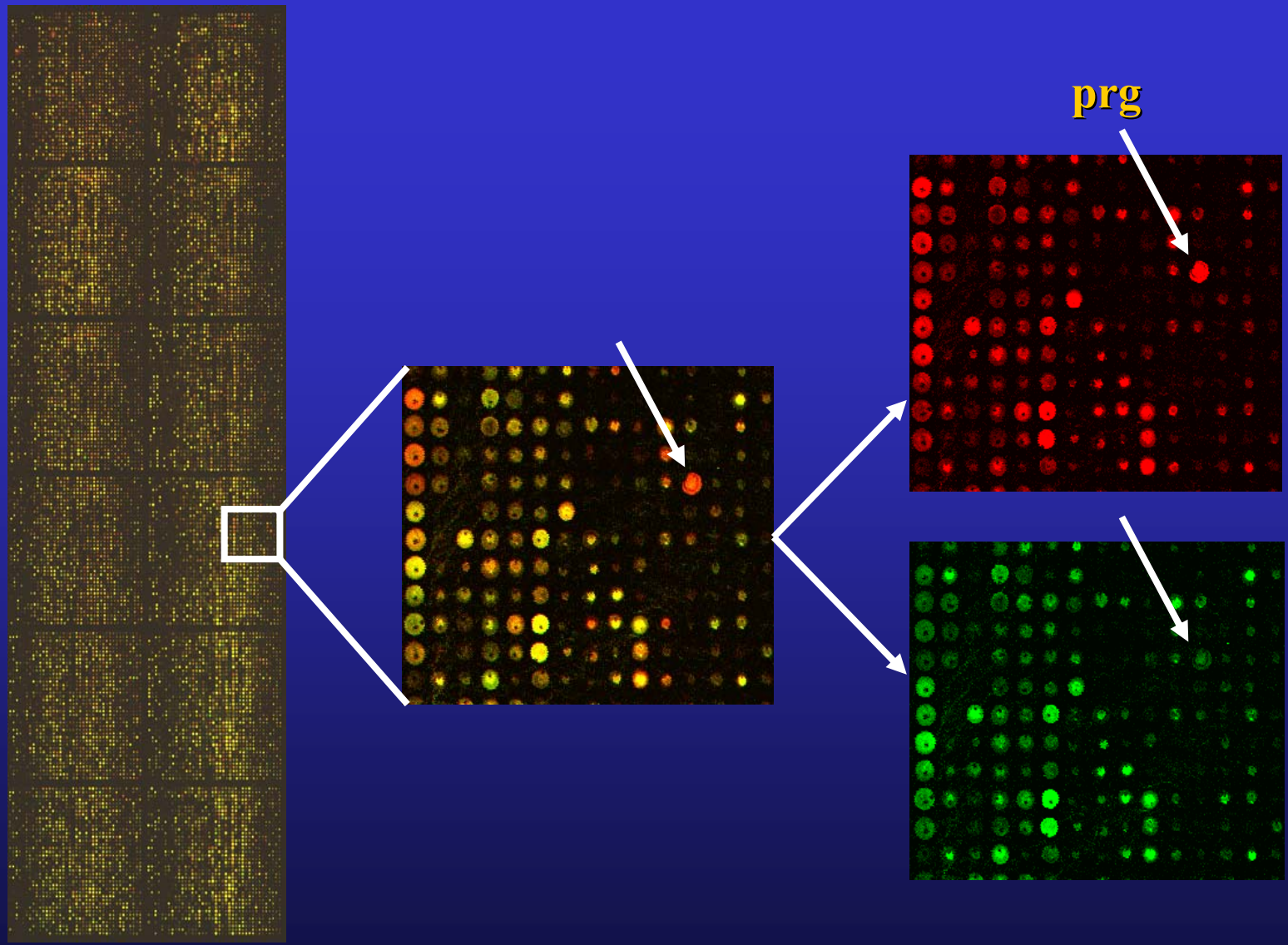
R/G ratio represents relative abundance of transcripts



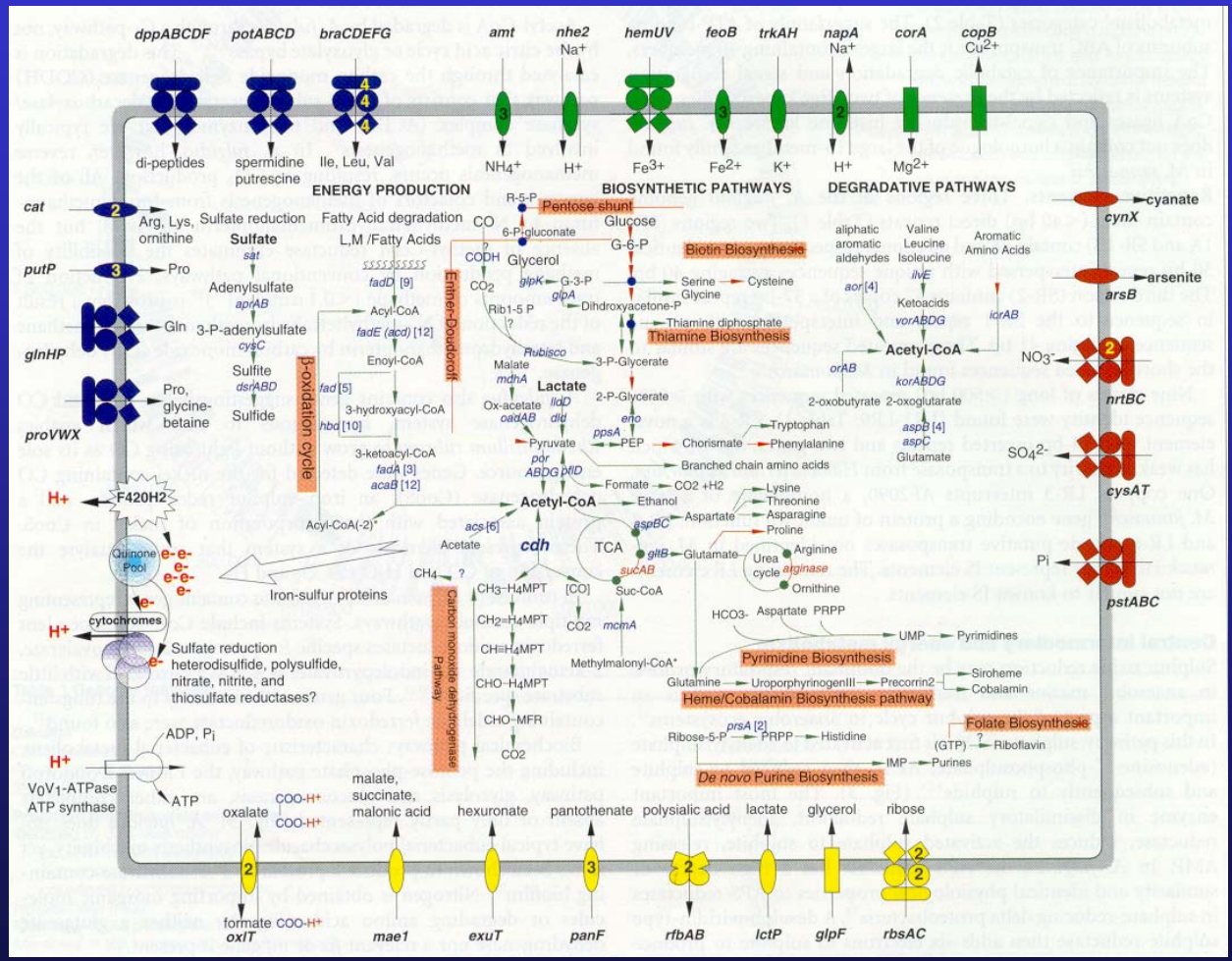
$- O_2$

$+ O_2$

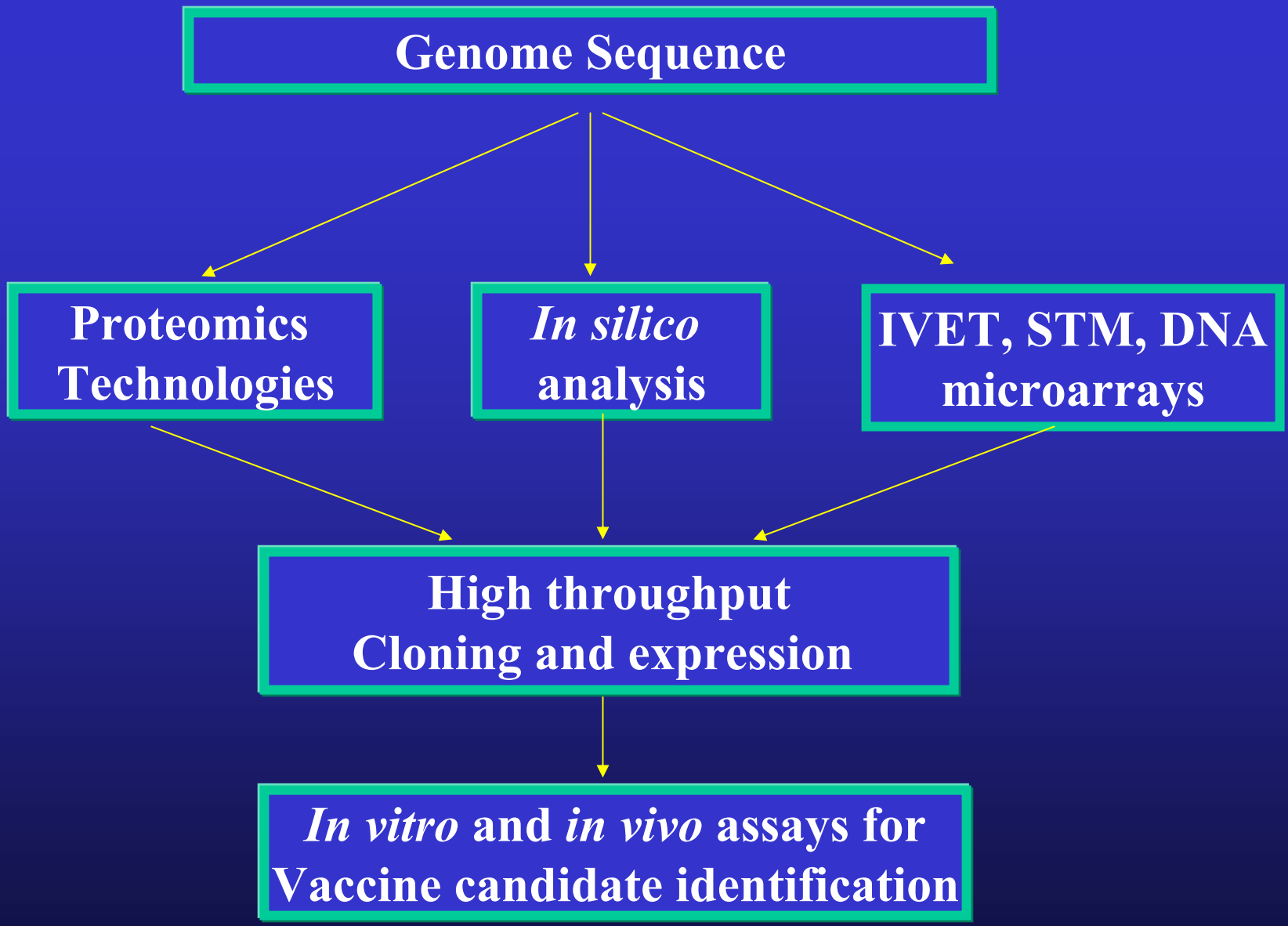
Gene Discovery: Identifying Novel Genes



In silico biochemistry

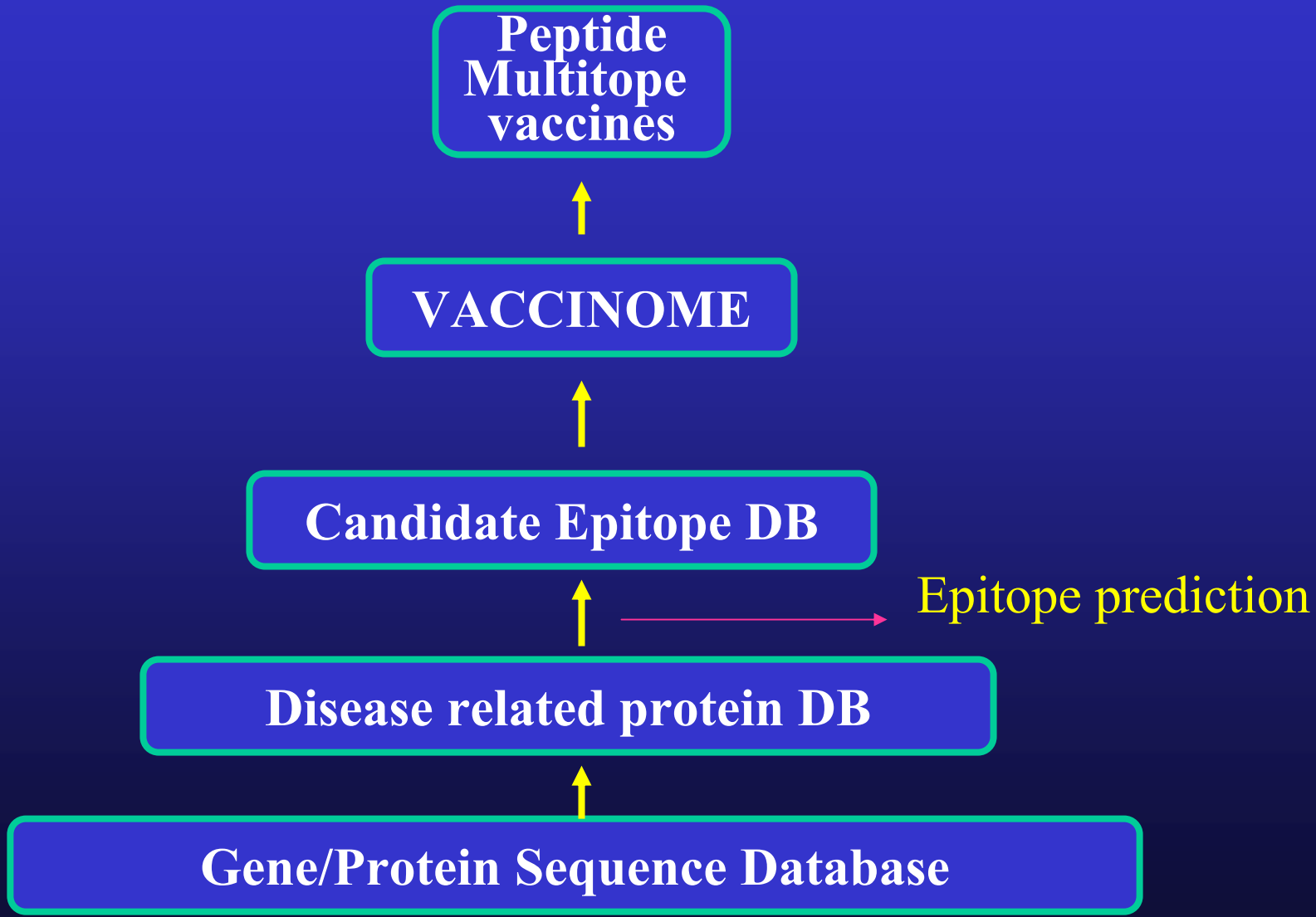


Metabolism and solute transport of *A. fulgidus*
Klenk *et al*, Nature n390, 364 (1997)



Global genomic approach to identify new vaccine candidates

In Silico Analysis

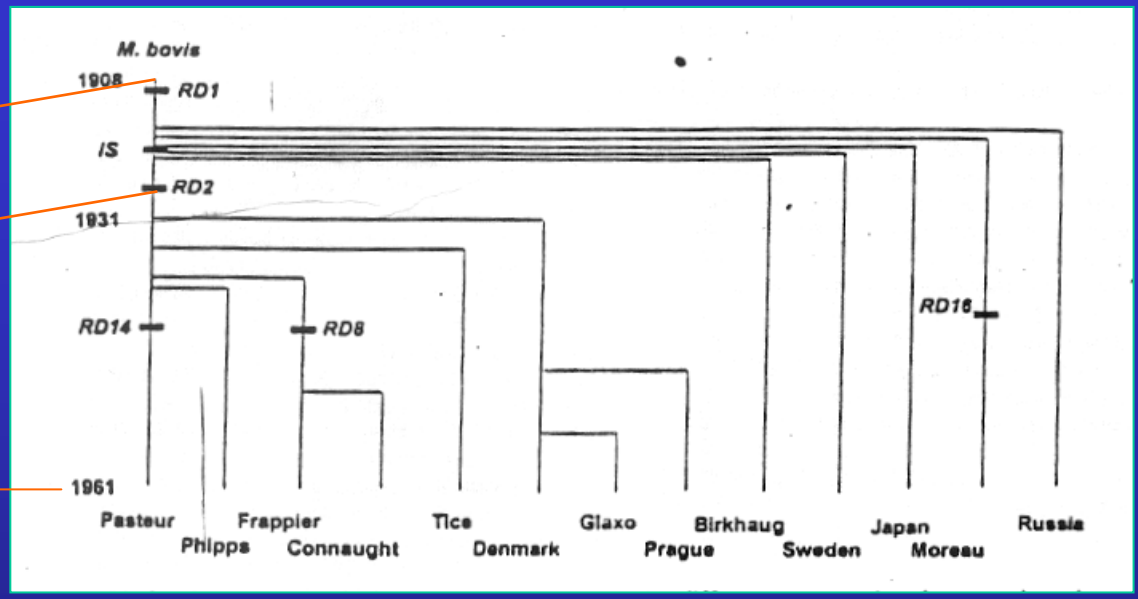


Comparative genomics of BCG vaccines by whole-genome DNA microarray

230 *in vitro* passages
(decreased virulence in animals)

(decreased vaccine lesions in humans,
+ decreased virulence in animals)

Lyophilized seed lots



Behr *et al.*, Science 284:1520, 1999

DNA microarray representing nearly all ORFs of *M. tuberculosis* H37Rv > parallel comparative hybridizations

Identified 16 regions deleted in BCG strains

9 RDs present in *M. tuberculosis* and absent from all *M. bovis* strains (including BCGs)

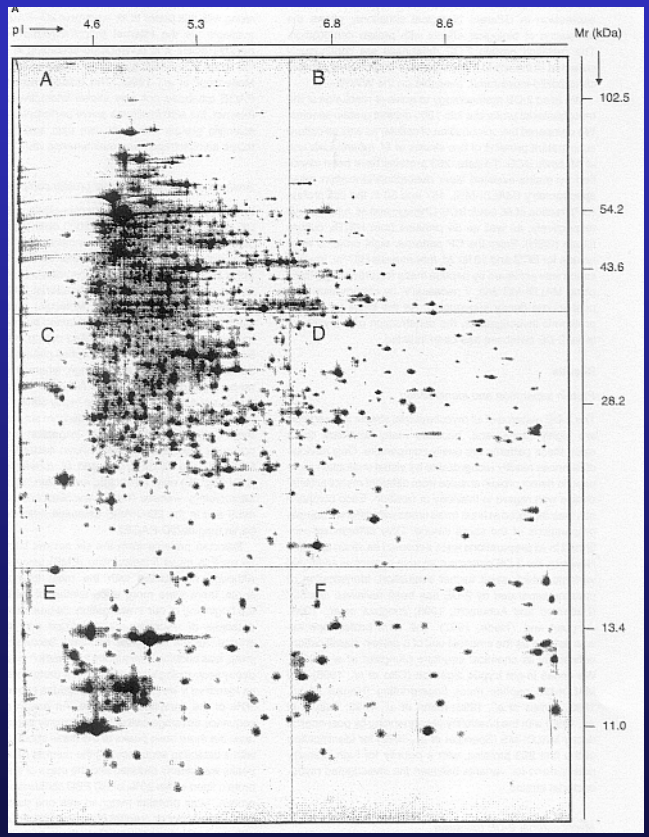
ORFs classified as transcriptional regulators are overrepresented in BCG deletions relative to their frequency in TB genome

« Deletions detected reflect a progressive adaptation of BCG strains to laboratory conditions that has compromised their capacity to survive within the host, impairing their ability to stimulate a durable immune response »

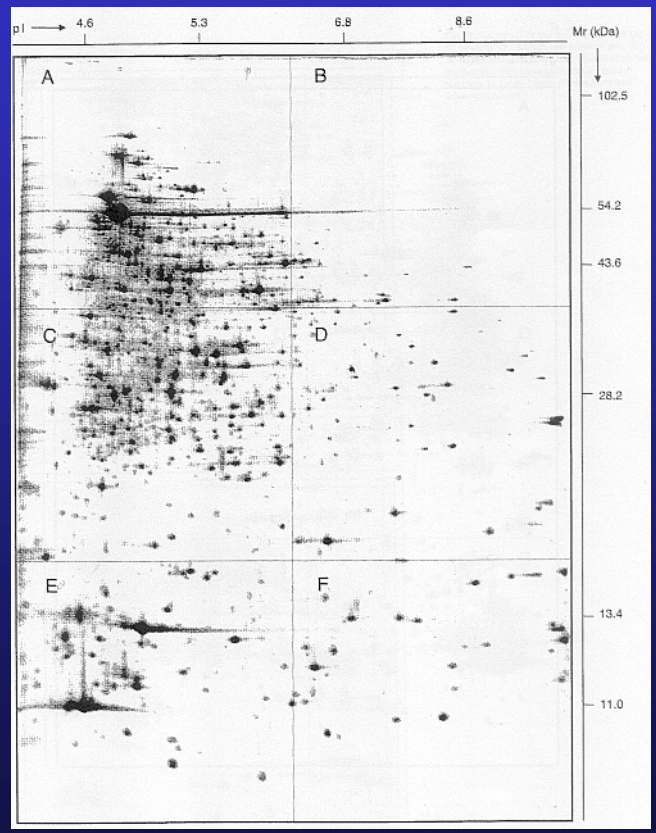
Molecular Microbiology (1999) 33(6), 1103–1117

Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens

Jungblut *et al.*, Max-Planck Inst. for Infection Biology

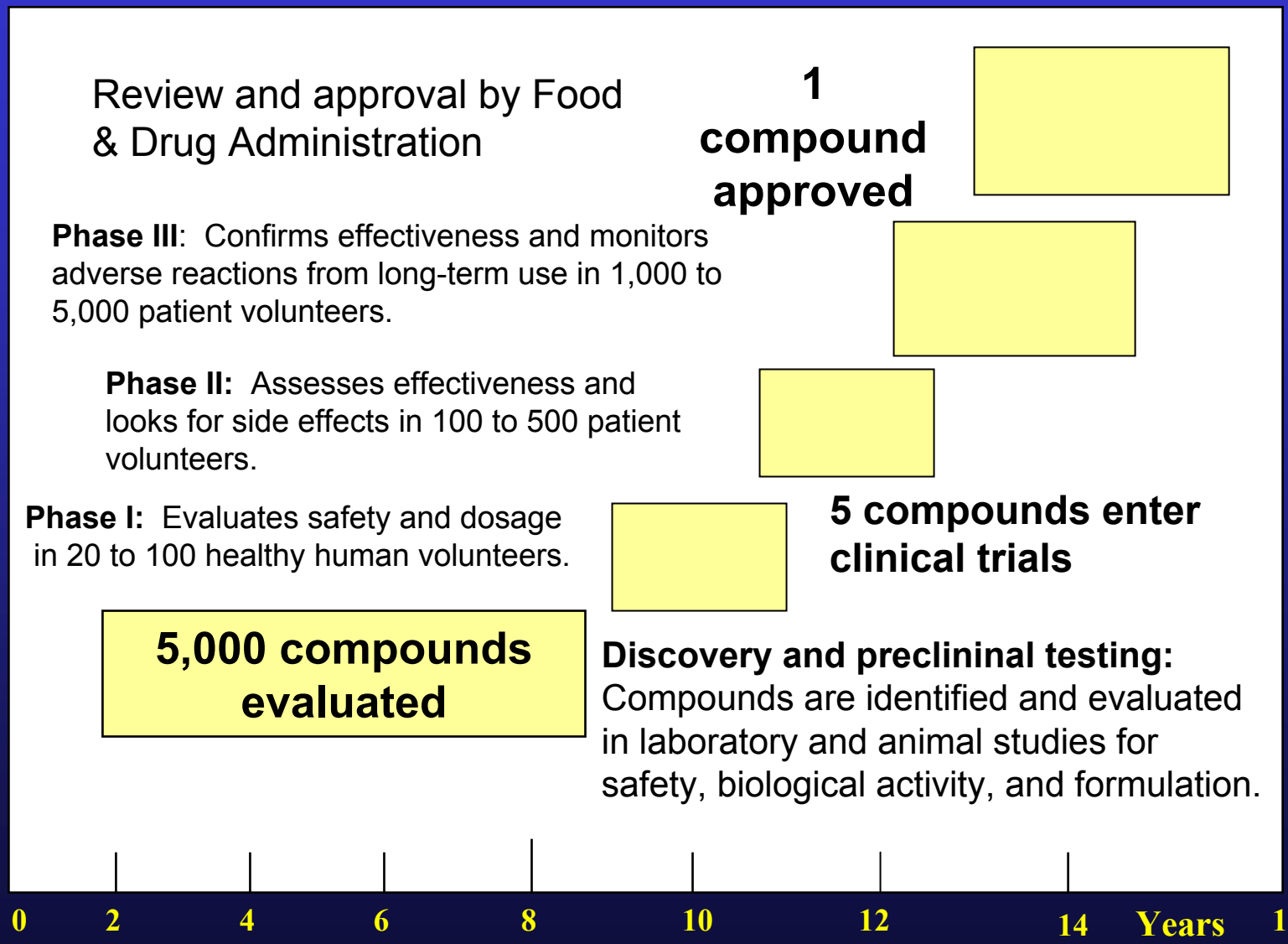


Mycobacterium bovis BCG



Mycobacterium tuberculosis H37Rv

Bringing a New Drug to Market



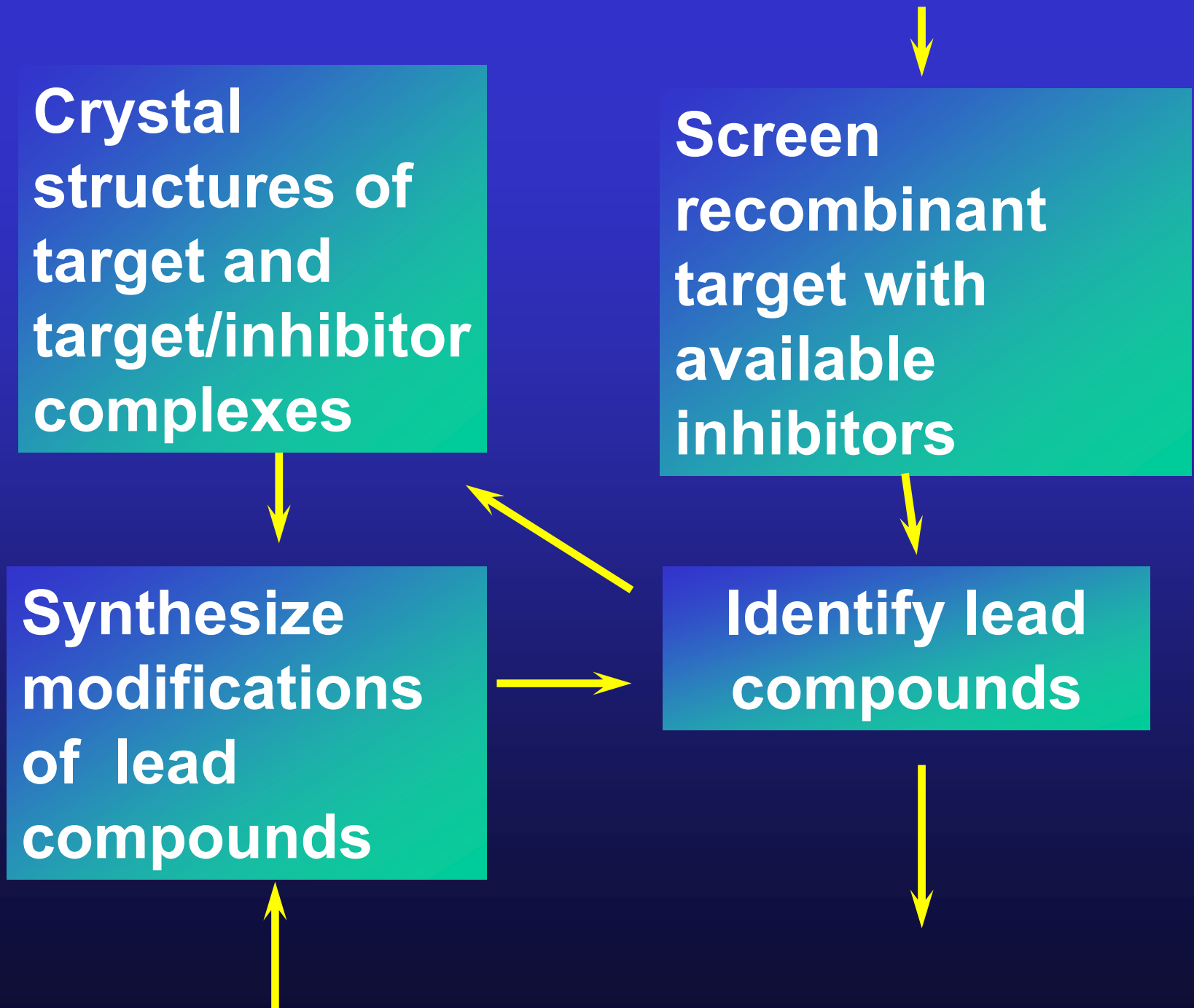
Source: Tufts Center for the Study of Drug Development

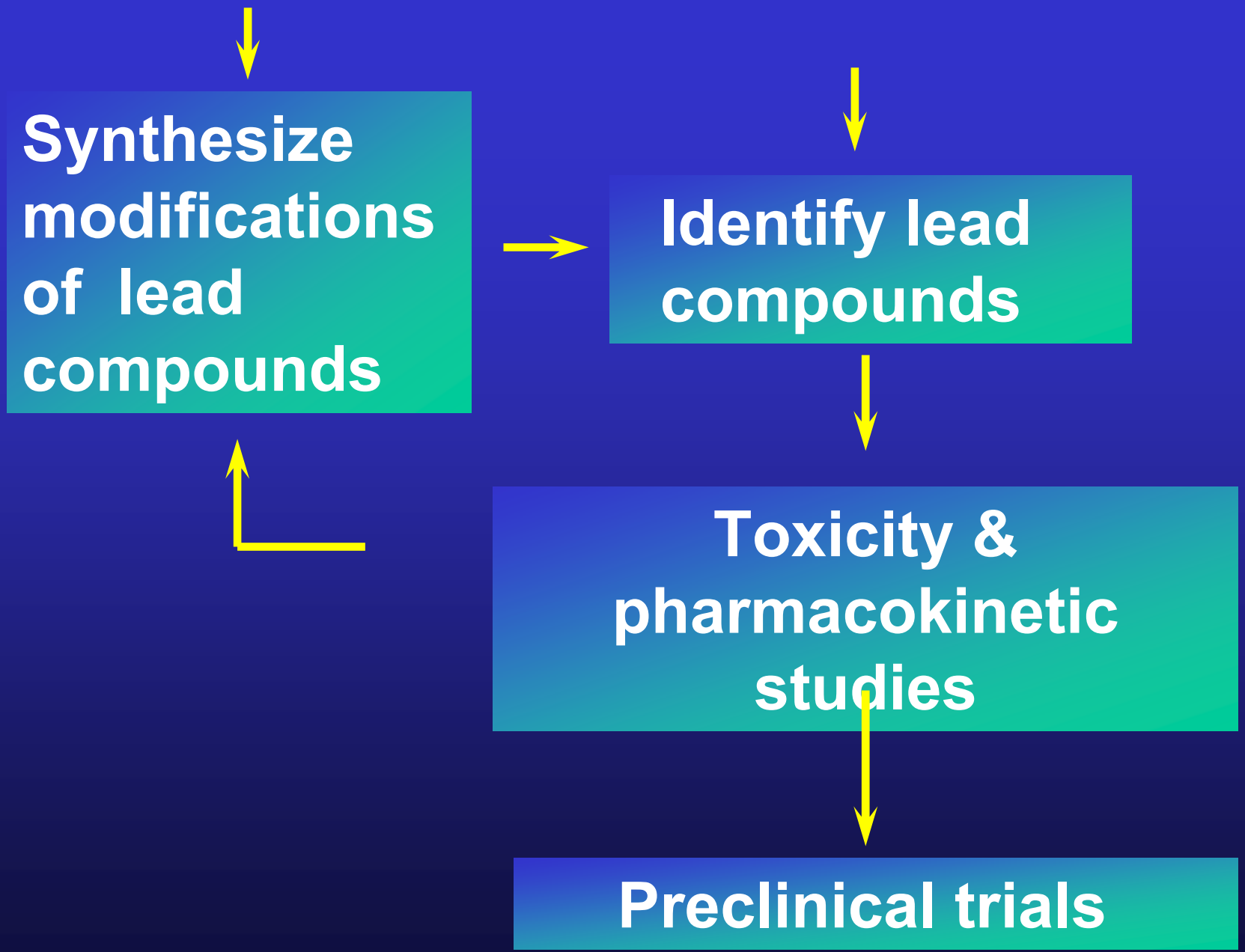
Rational Approach to Drug Discovery

Identify target

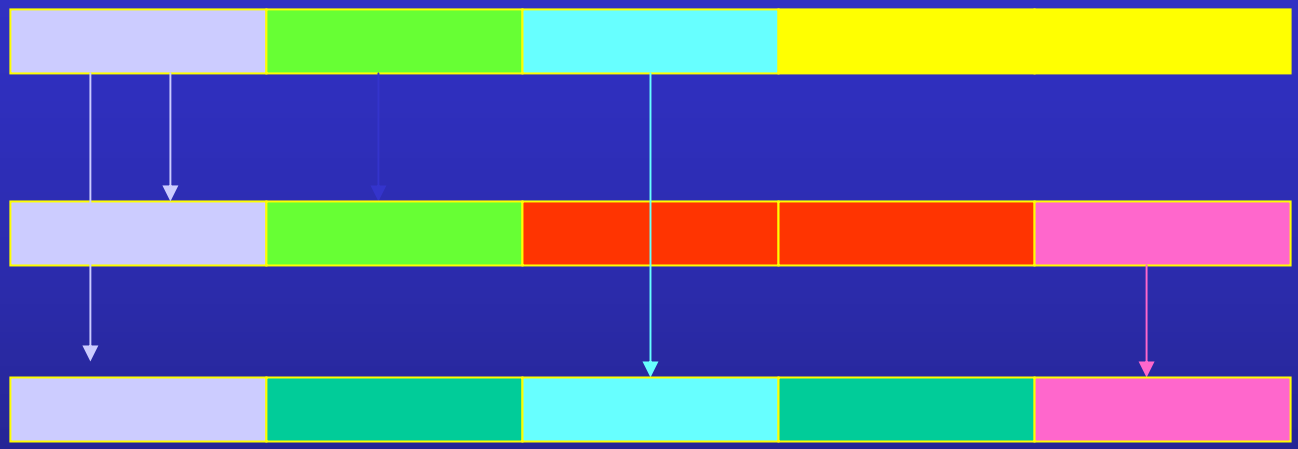
Clone gene encoding target

Express target in recombinant form


Crystal structures of target and target/inhibitor complexes
Screen recombinant target with available inhibitors
Synthesize modifications of lead compounds
Identify lead compounds


Synthesize modifications of lead compounds
Identify lead compounds
Toxicity & pharmacokinetic studies
Preclinical trials

What one should look for?



Proteins that are shared by –

- **All genomes**
- **Exclusively by Human & P.f.**
- **Exclusively by Human & Mosquito**
- **Exclusively by P.f. & Mosquito**

Unique proteins in –

Human

P.f. Targets for anti-malarial drugs

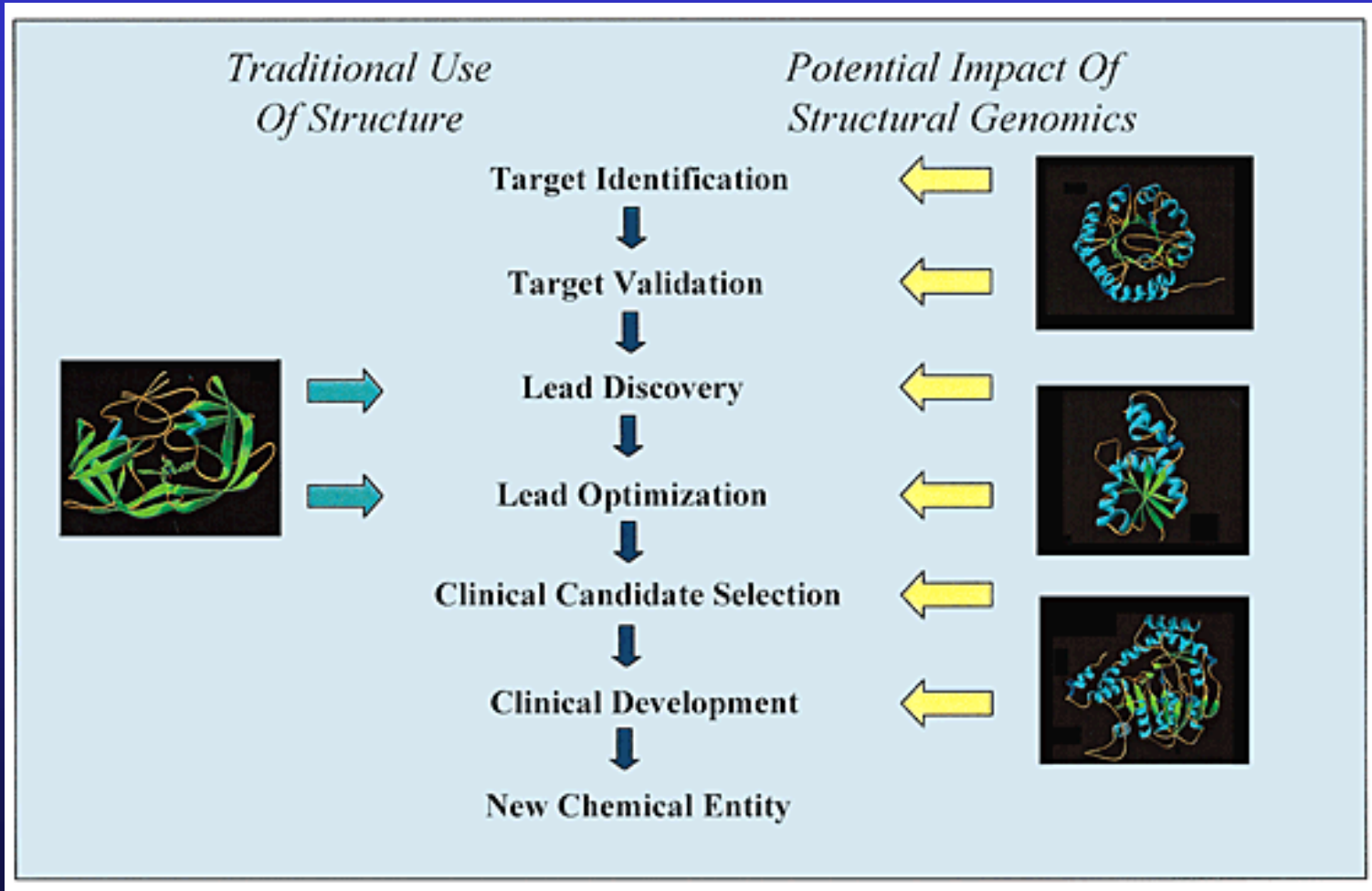
Mosquito

How Bioinformatics can help in Target Identification?

- Homologous & Orthologous genes
- Gene Order
- Gene Clusters
- Molecular Pathways & Wire diagrams
- Gene Ontology

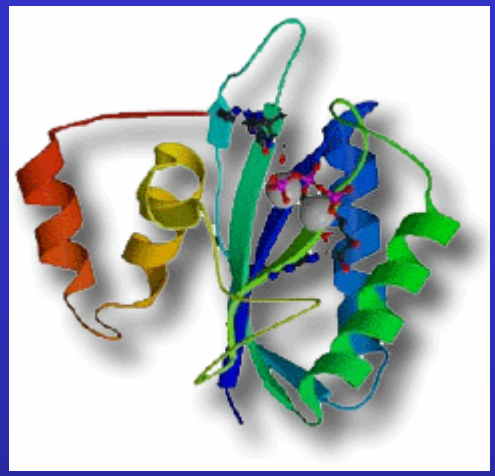
Identification of Unique Genes of Parasite as potential drug target.

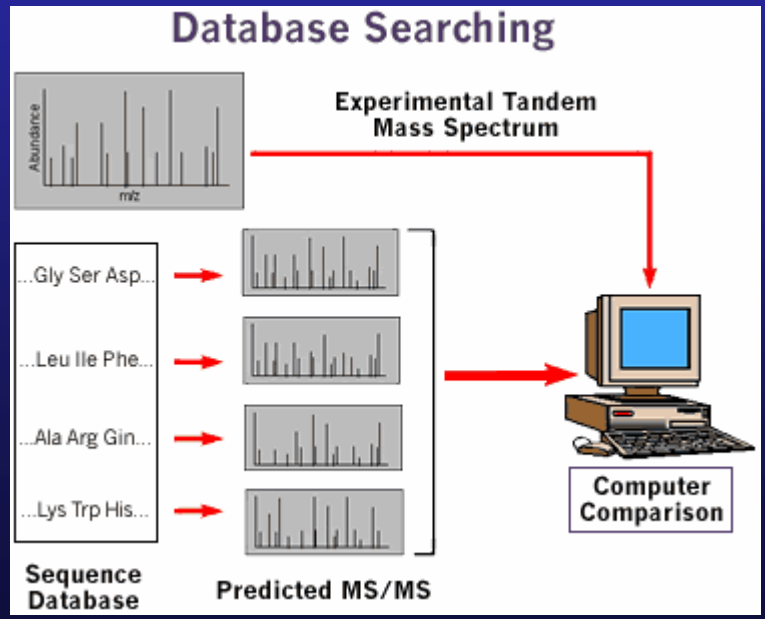
Impact of Structural Genomics on Drug Discovery

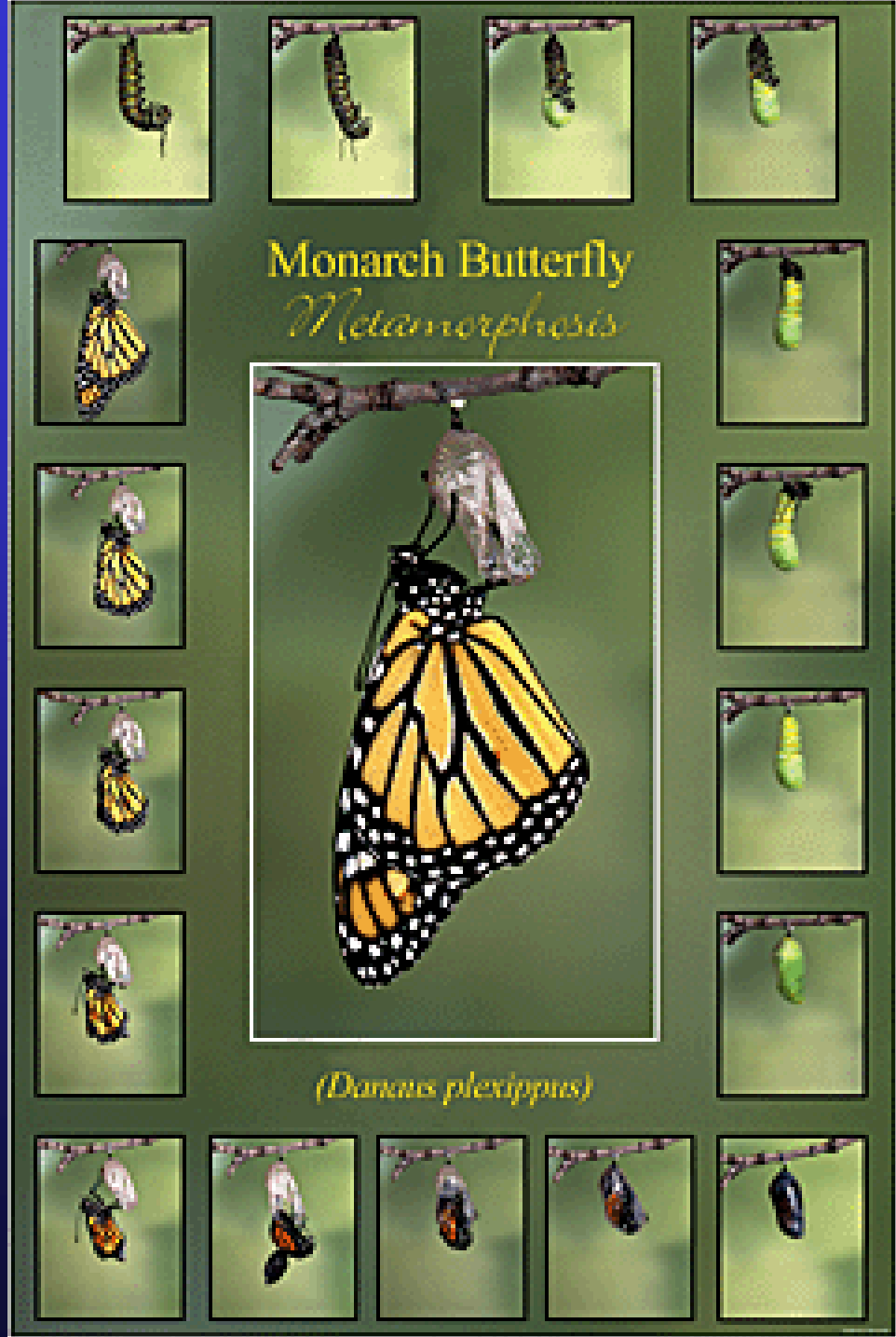


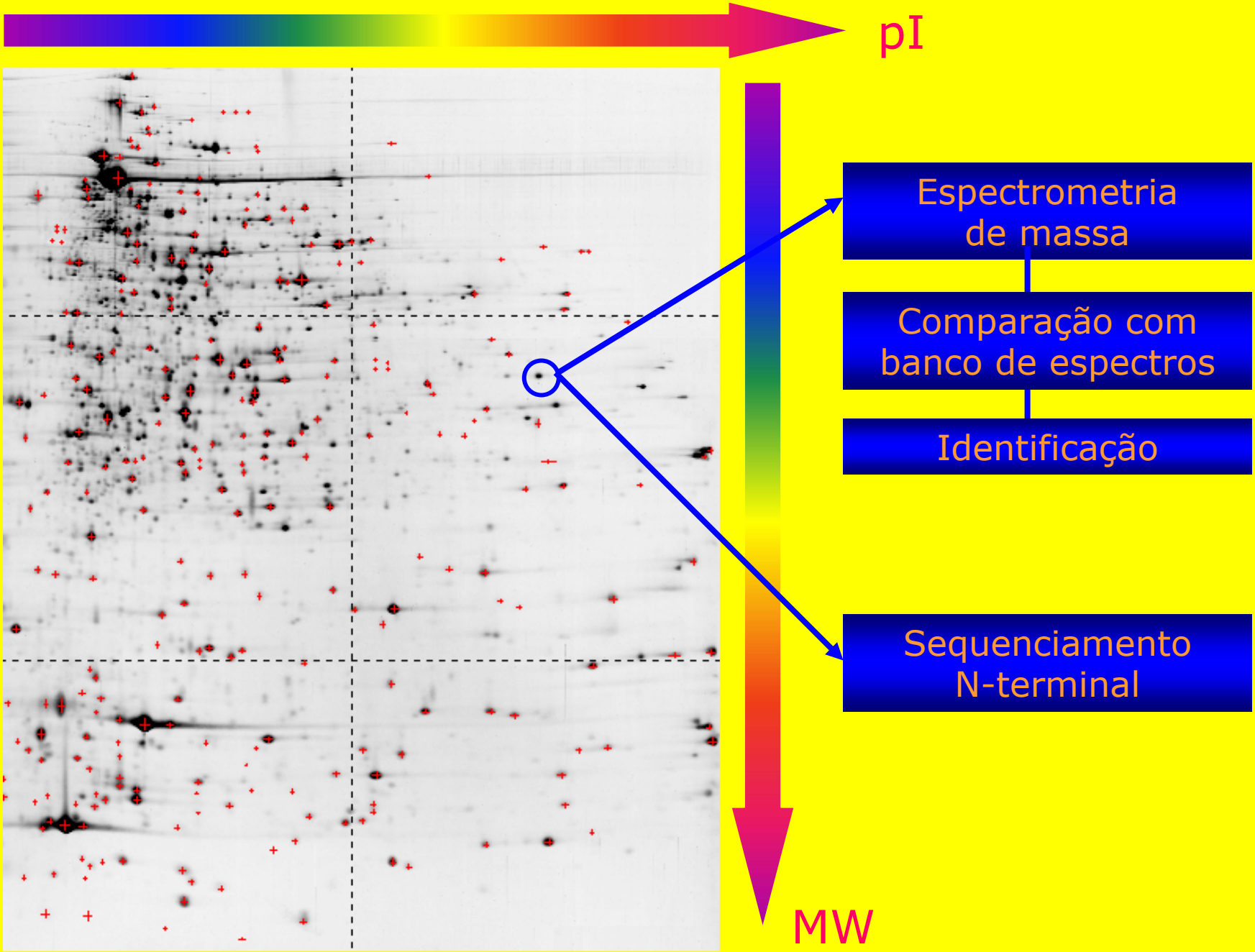
Dry, S. et. al. (2000) Nat. Struc.Biol. 7:976-949.

Proteômica




Database Searching
Abundance
m/z
Experimental Tandem Mass Spectrum
...Gly Ser Asp...
...Leu Ile Phe...
...Ala Arg Gin...
...Lys Trp His...
Computer Comparison
Sequence Database
Predicted MS/MS

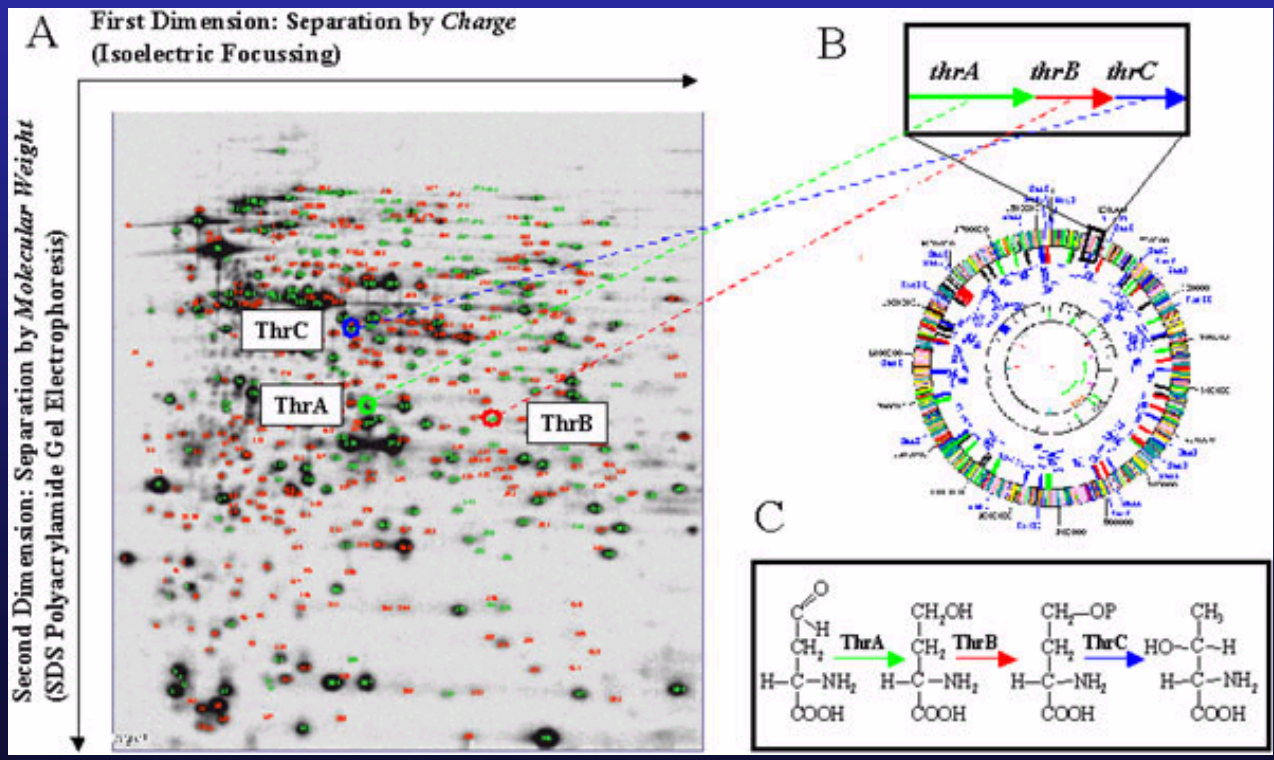

Monarch Butterfly
Metamorphosis
(Danaus plexippus)


pI
MW
Espectrometria de massa
Comparação com banco de espectros
Identificação
Sequenciamento N-terminal

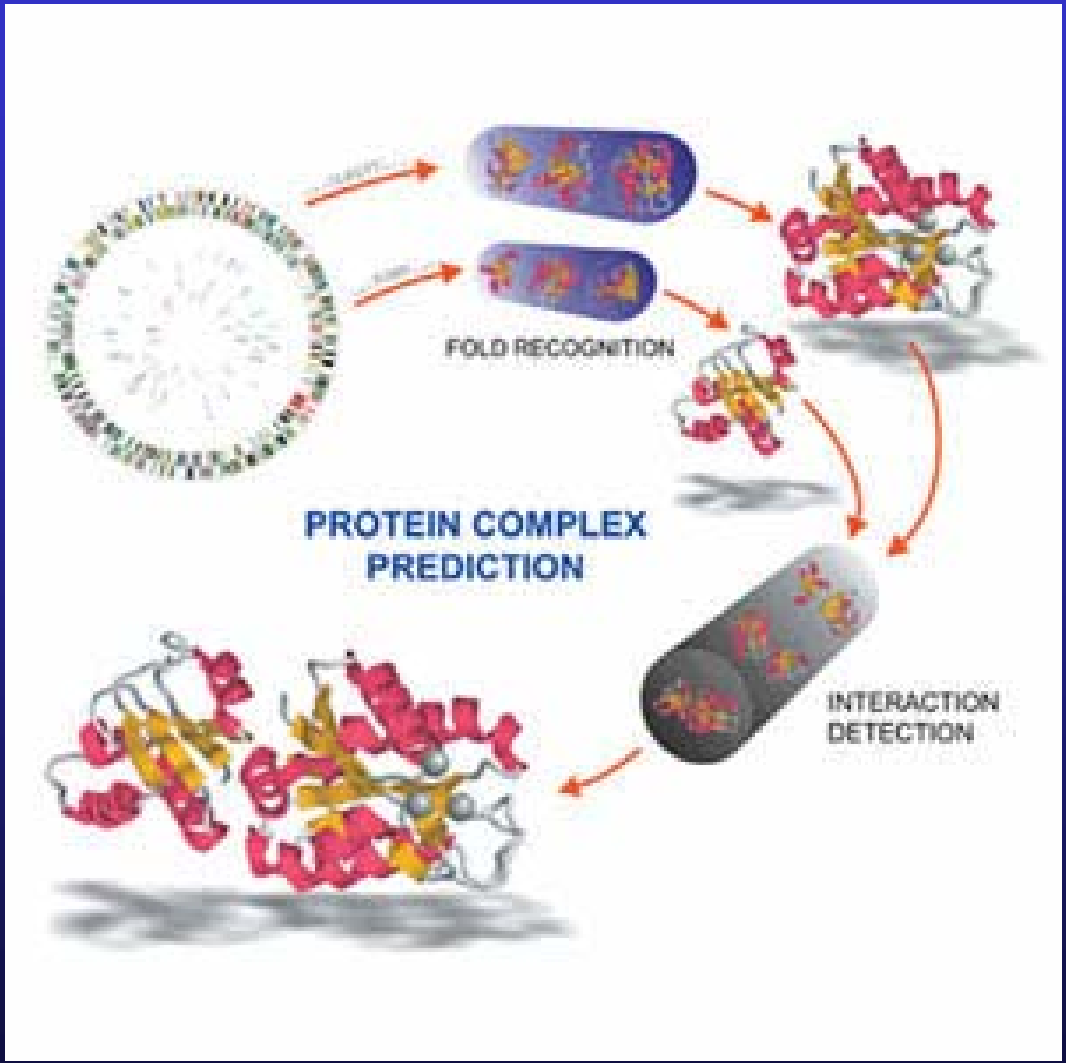
Genômica funcional: o *metaboloma*

O terceiro nível de análise funcional = Metaboloma

Complemento de todos os metabólitos de baixo peso molecular de uma suspensão celular (ou amostra) de interesse. Mudanças na concentração de enzimas individuais tendem a ter pouco efeito em fluxos metabólicos específicos (ou no fenótipo observado sob condições de laboratório). Entretanto, mudanças na concentração de enzimas específicas podem e têm um efeito substancial nas concentrações de metabólitos.



Interações proteína - proteína:



“Os dois desafios principais do novo milênio são:

1. O gerenciamento de informação excessiva: Como fazer genômica?
2. A excessiva complexidade dos sistemas.

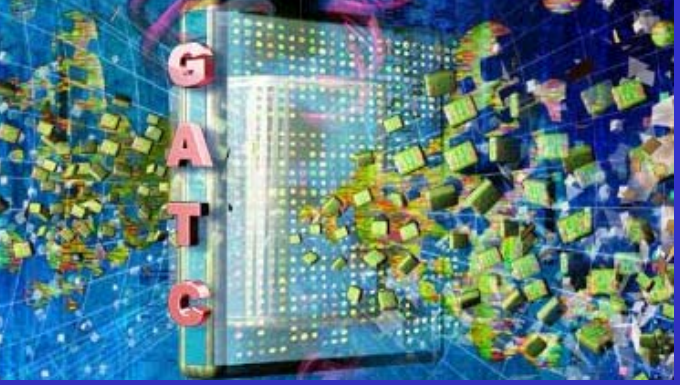
- Decifrar em detalhe as interações moleculares entre as moléculas efetoras de um patógeno e seus alvos celulares
- Integrar a microbiologia celular no esquema mais amplo da infecção do hospedeiro
- Abordar os conceitos de resistência e sensibilidade do hospedeiro à infecções e a imunoreatividade de patógenos, que podem influenciar o grau de severidade de uma doença infecciosa
- Identificar os fatores necessários à sobrevivência de patógenos no meio ambiente (‘vironmence’ factors vs. ‘virulence’ factors).”

Philippe J. Sansonetti

Trends in Microbiology, **8**: 196-197, 2000.

"... and that's what these young men don't always realize, that you've got to learn a lot of hard thinking in order to have bright ideas, you see."

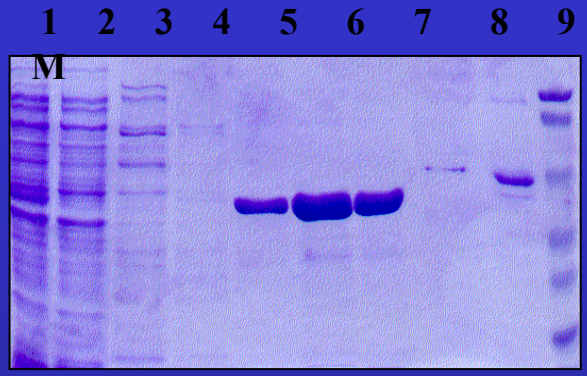
Francis Crick



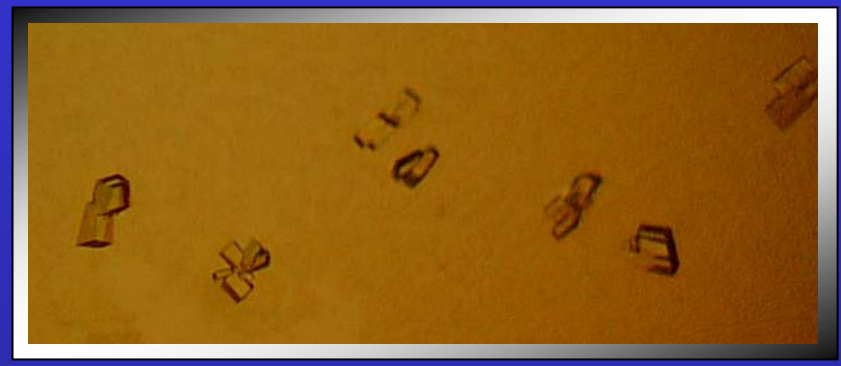
Links:

A Science Primer:
http://www.ncbi.nlm.nih.gov/About/primer/index.html
Access Excellence:
http://www.accessexcellence.org/
DNA Interactive:
http://www.dnai.org/index.html
DNA from the Beginning:
http://www.dnaftb.org/dnaftb/
Genome News Network:
http://gnn.tigr.org/timeline/timeline_home.shtml
New York Times 'DNA at 50':
http://www.nytimes.com/indexes/2003/02/25/health/genetics/index.html
Microbiology's Most Significant Events:
http://www.microbeworld.org/htm/aboutmicro/timeline/tmln_0.htm

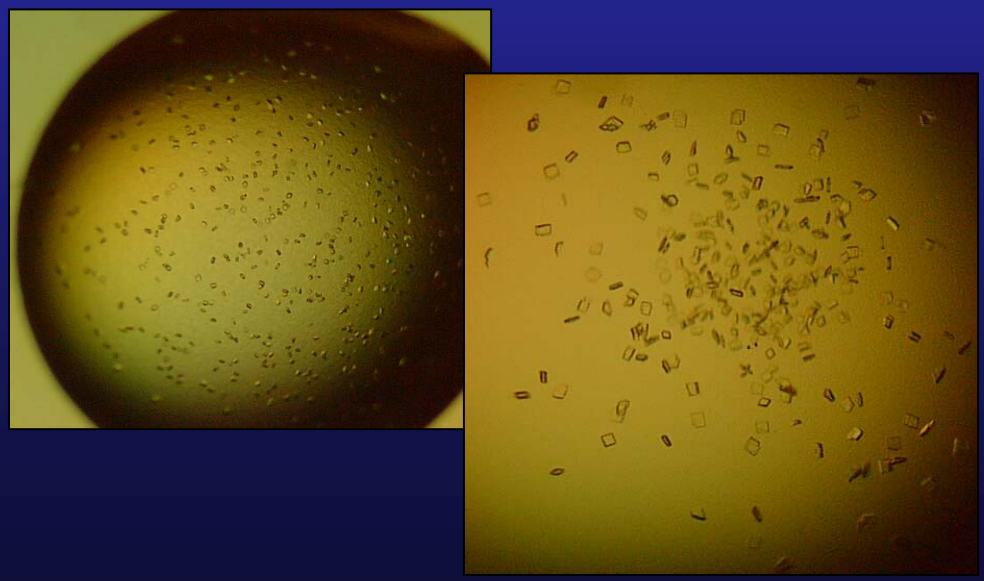
Structural analysis of rTc45



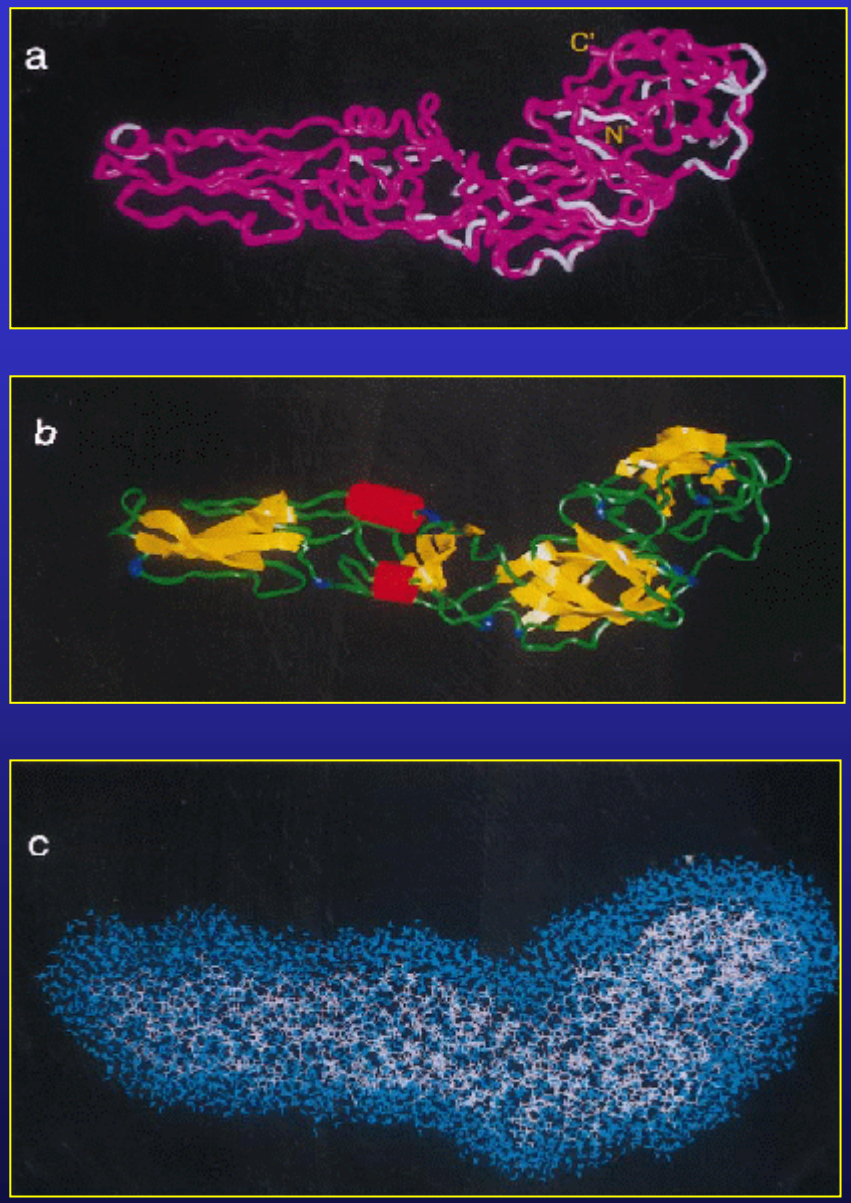
(**1**) *E. coli* B1 total cleared extract , (**2**) flow-through IMAC collumn, (**3**) 30mM and (**4**) 60mM imidazol washes, (**5-7**) eluted fractions, (**9**) rTc45 B6 protein



rTc45-A after IMAC purification (resolution 7 A)



FPLC purified rTc45-A (resolution 2.8 A)


a
C'
N'
b
c

FIM