Biologia Computacional
Um Espaço Desafiador para os Profissionais de Computação

Katia S. Guimarães
katiag@cin.ufpe.br

Seminário CIn – 14/dezembro/2007

Roteiro

- Breve Histórico da Era **Genômica**
- Pós-Genoma, a Curta Era **Proteômica**
- A Essencialidade do **Interactoma**
- Nosso Trabalho de Pesquisa e Oportunidades no CIn

Bio-informática *vs.* Biologia Computacional

Há muita controvérsia quanto a estes termos.

Para alguns

Bio-informática é uma especialização da Informática que trata de desenvolver ferramentas para lidar com dados biológicos.

Biologia (Molecular) Computacional área de pesquisa que combina conhecimentos de Química, Física, Biologia, C. Computação, Matemática e Estatística para atacar problemas de Biologia Molecular.

Histórico da Era Genômica - 1990

Início: Outubro de 1990

Lançamento do Projeto Genoma Humano

- Seqüenciar o DNA humano ($3 \cdot 10^9$ pb) e
- Identificar os estimados 100 mil genes.

Atores Principais:

Consórcio envolvendo EUA, Inglaterra, França, Japão, Alemanha e China.

Prazo: 15 anos (terminaria em 2005)

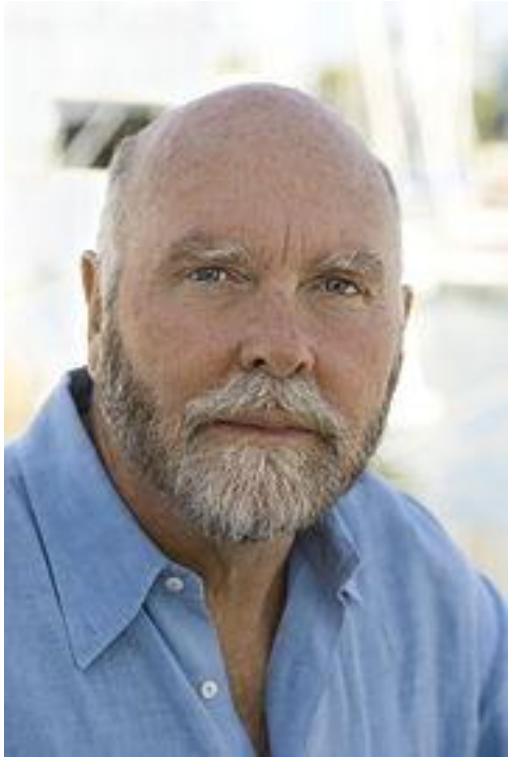
Orçamento: US$3 bilhões de dólares

Histórico da Era Genômica - 1992

1992

→ Consórcio faz mapas dos cromossomos humanos

→ Craig Venter, pesquisador do NIH, funda *The Institute for Genomic Research* (TIGR)



Histórico da Era Genômica - 1995

Grupo de pesquisadores da TIGR publica na revista *Science* o artigo

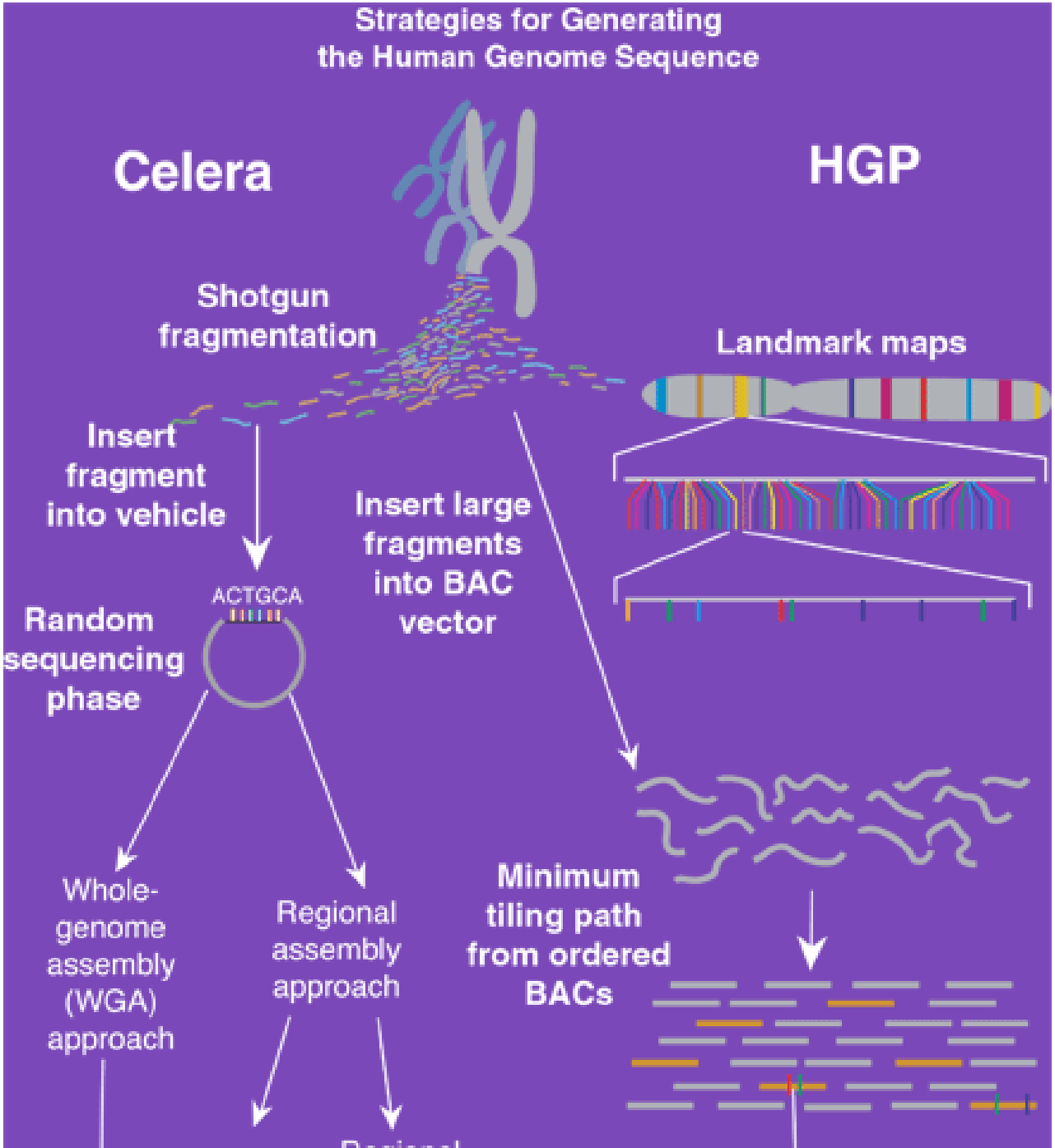
Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.
Fleischmann RD, Adams MD, et al.

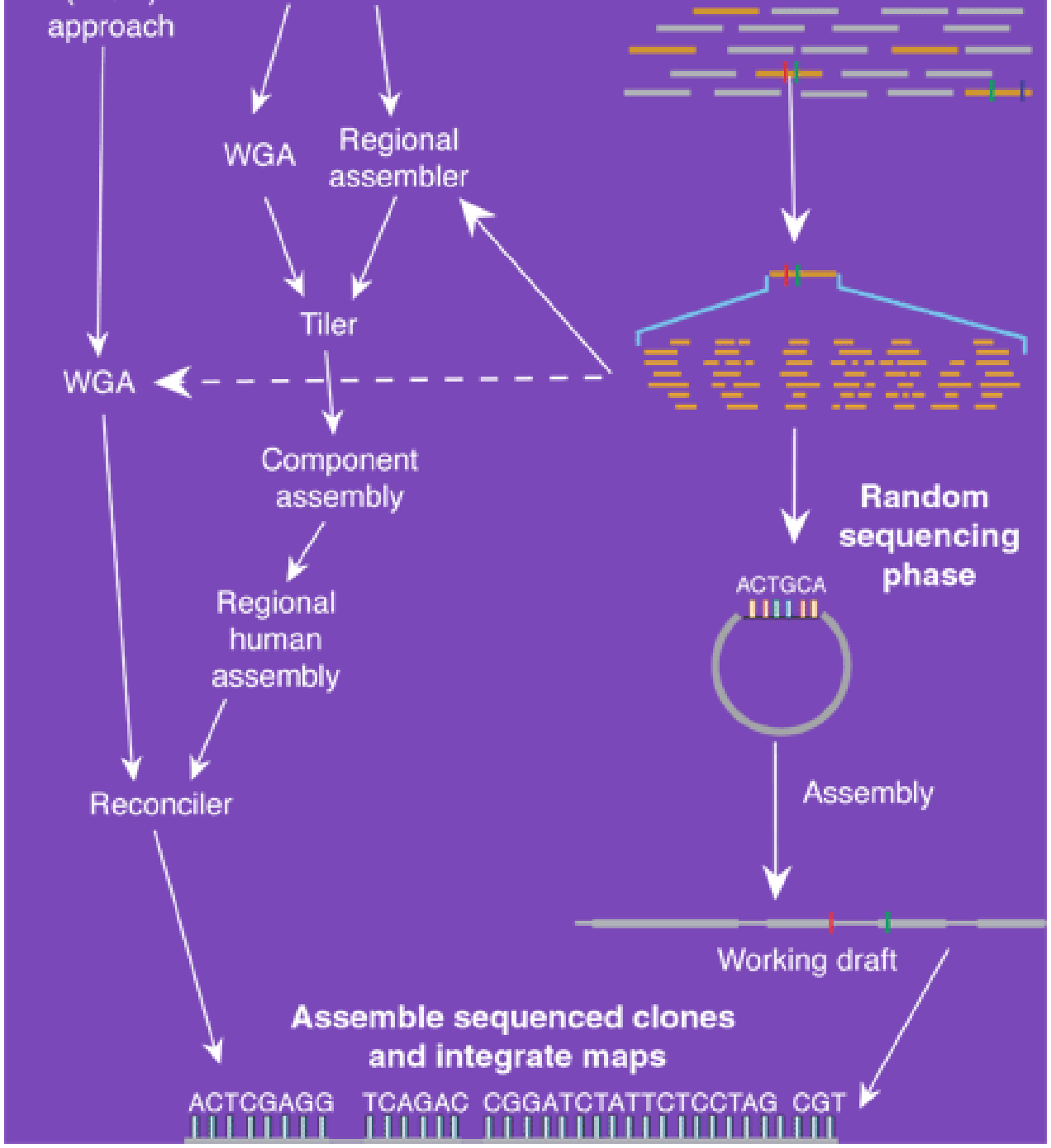
com a seqüência de DNA da bactéria *Haemophilus influenzae (otite, meningite)*

Tamanho: $2 \cdot 10^6$ bp

Técnica: *double-barrel shotgun sequencing*

- Mais custo computacional
- Muito menos tempo e custo em labs.


Strategies for Generating
the Human Genome Sequence
Celera
HGP
Shotgun
fragmentation
Landmark maps
Insert
fragment
into vehicle
Insert large
fragments
into BAC
vector
ACTGCA
Random
sequencing
phase
Whole-
genome
assembly
(WGA)
approach
Regional
assembly
approach
Minimum
tiling path
from ordered
BACs


approach
WGA
Regional assembler
Tiler
WGA
Component assembly
Regional human assembly
Reconciler
Random sequencing phase
ACTGCA
Assembly
Working draft
Assemble sequenced clones and integrate maps
ACTCGAGG
TCAGAC
CGGATCTATTCTCCTAG
CGT

Histórico da Era Genômica – 96-98

Pesquisadores da TIGR publicam as seqüências de DNA de outras bactérias

Mycoplama genitaluim (1996) (menor bactéria)

Methanococcus jannaschii (1997)

1998

Craig Venter se associa com a Applied Biosystems para fundar a *Celera Genomics Corp., com o objetivo de seqüenciar o genoma humano em 03 anos (2001),*

ao custo de US$300 milhões (1/10 do orçamento do projeto do Consórcio).

Histórico da Era Genômica – 99-00

O Consórcio revê as suas previsões, e anuncia a conclusão do seqüenciamento para 2003,

.

D

amento do primeiro

cromossomo

consórcio mundial.

Ju

unciam a conclusão

enoma humano.

Histórico da Era Genômica – Fev 2001

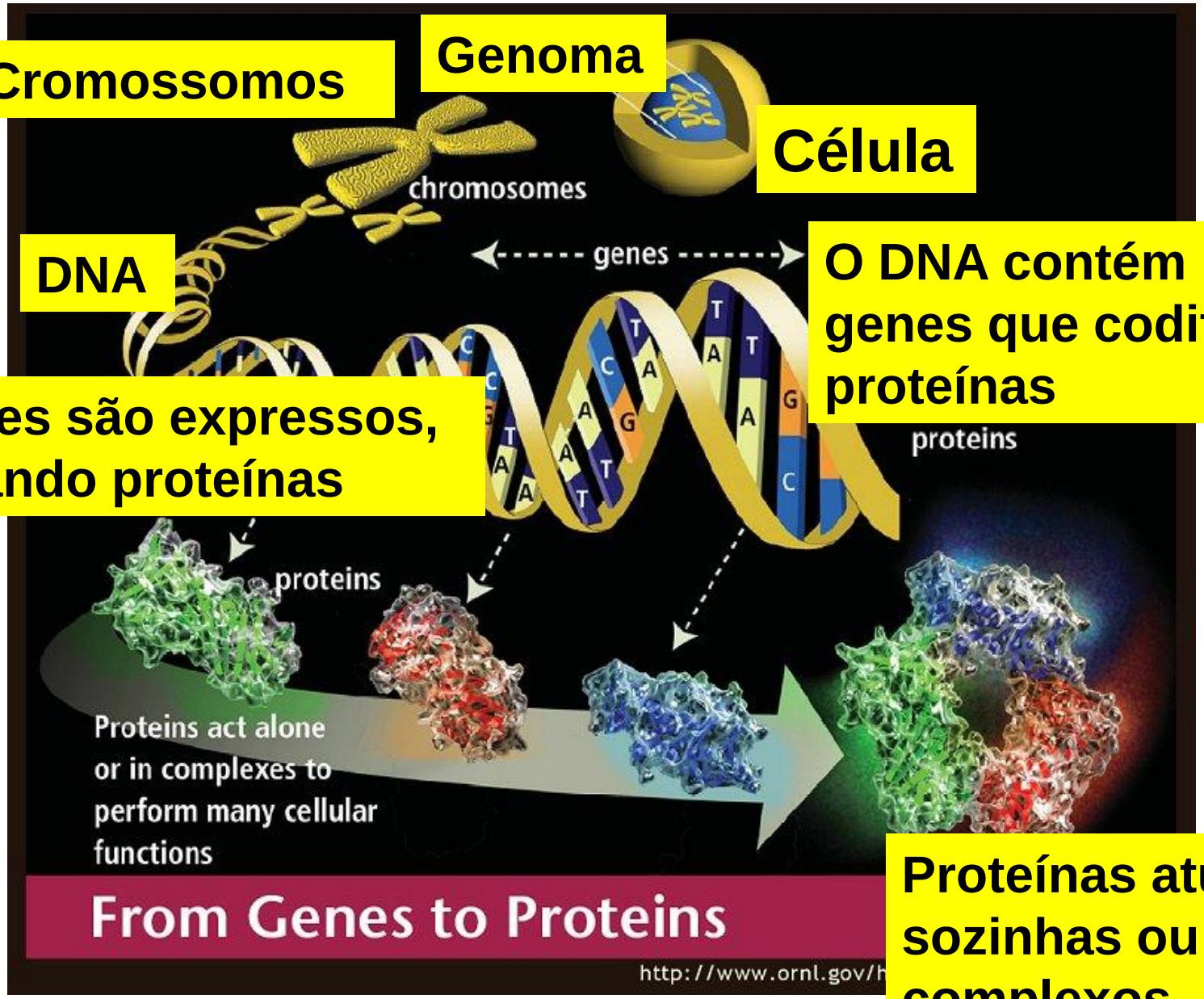
The Sequence of the Human Genome

J. C. Venter, M. D. Adams, E. W. Myers, et al.

Initial Sequencing and Analysis of the Human Genome

Consórcio Mundial

Biologia Molecular 101 em 2 Minutos



Cromossomos

Genoma

Célula

DNA

O DNA contém genes que codificam proteínas

Genes são expressos, gerando proteínas

Proteínas atuam sozinhas ou em complexos

O GenBank

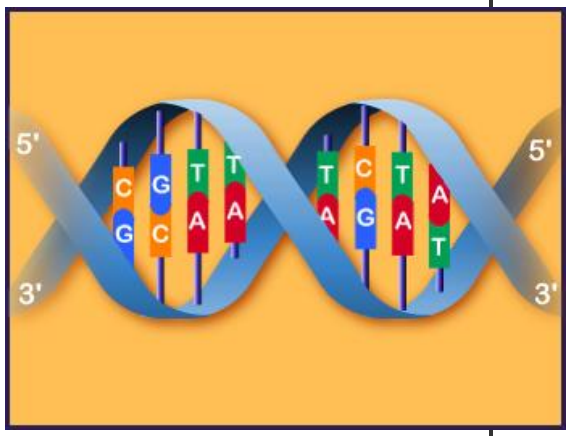
Do ponto de vista da computação, foram criadas inúmeras seqüências sobre o alfabeto {a, c, g, t}

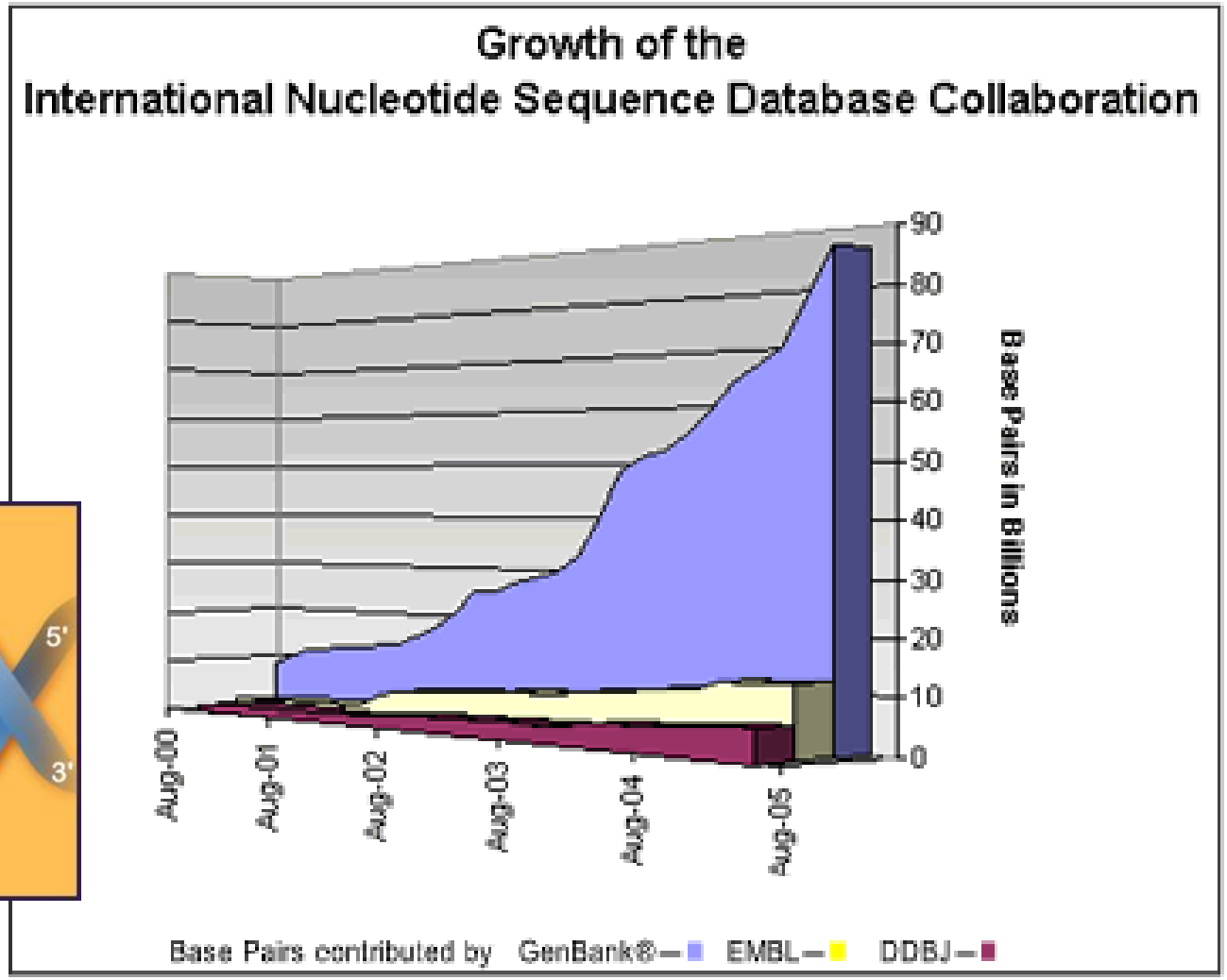
```
   1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
  61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
 121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
 181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
 241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
 301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
 361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
 421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
 481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
 541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
 601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
 661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
 721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
 781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
 841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
 901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
 961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
```

Link: http://www.ncbi.nlm.nih.gov/Genbank/index.html

O GenBank do NCBI

Aproxima-damente 85 bilhões de pares de base





Link: http://www.ncbi.nlm.nih.gov/Genbank/index.html

O que fazer com tanta Letrinha?

O genoma é como um livro

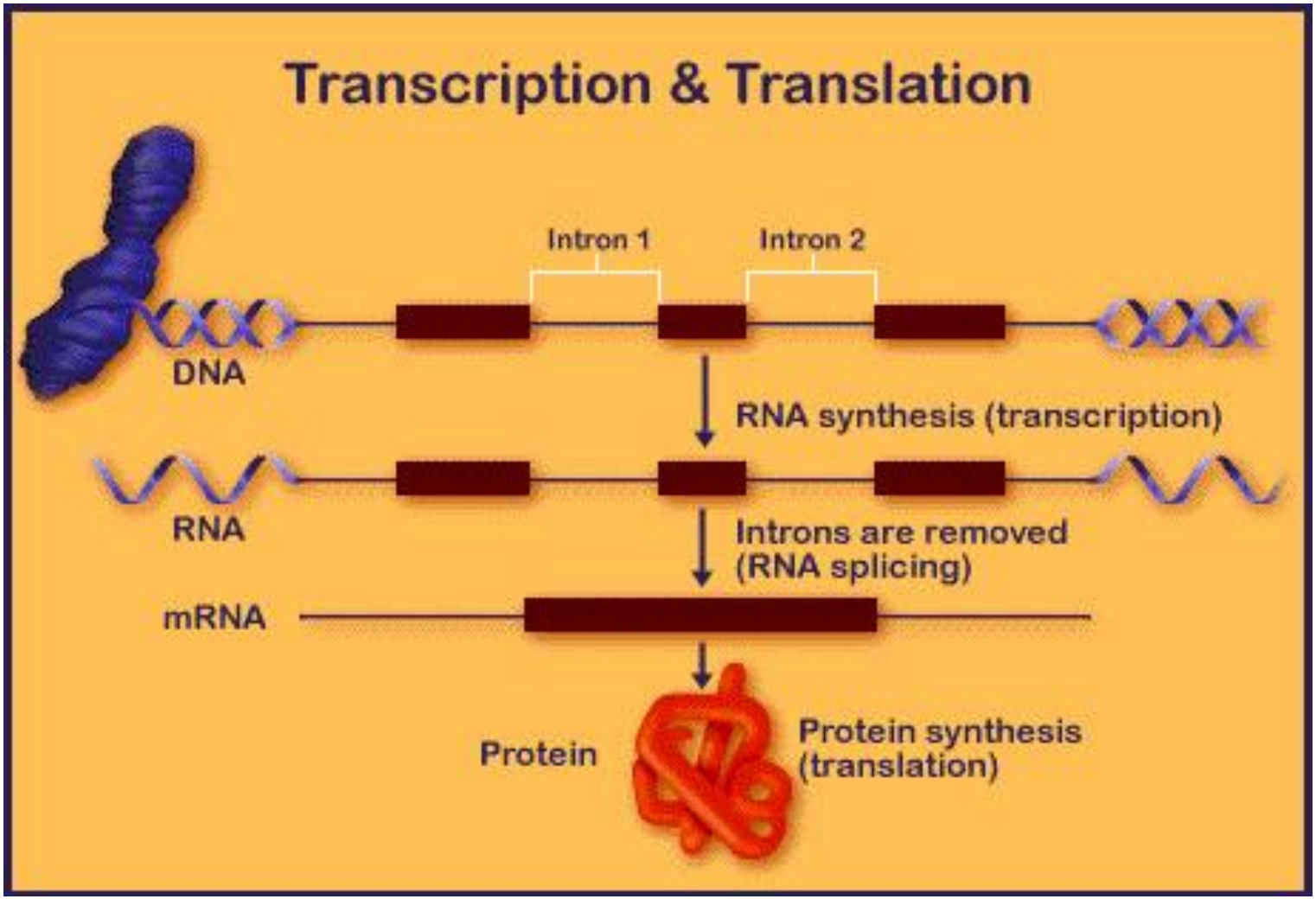
dividido em capítulos (cromos-somos)

e estes em palavras (genes)

Rat prion gene D50092 upstream of exon 1: annotations

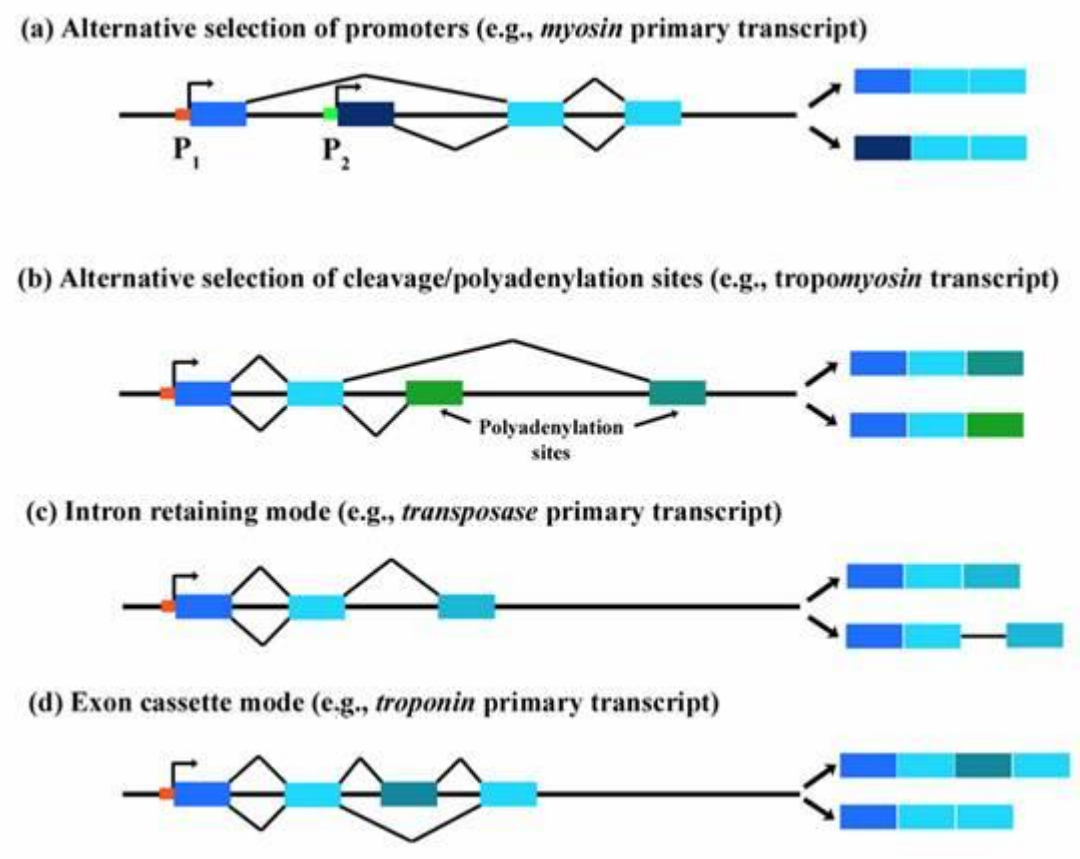
```
   1 aagcttttaa agccacttcc tggtgggaag agagcagtca gccagtaggt atttgattct
  61 tagaggaaaa agctgatatc ttttaaccaa gcctttcaga gtccccctgt gggagaggcc
 121 agatggaggt gaggggaagg ctggtccctt ccctaagtga atctacccac tcctctcctg
 181 cgtcccctgc catctctgac agcaagtgga gggcctctga gaaattcaaa gatgcggtca  pre-gene region
 241 ccctgatcac aatgagtact ctctgaggca ggaaggcaag gatctgaaag acggtcagtg
 301 tattccttaa gaacacagag taaatgttaa gcagcaccag caccactcat tctaacaggc
 361 tgaaatctgg gtcatctcag ccaaaggctc tgctccagcc ctctcatgca tgtcctcttc
 421 ccaaacagct gtcaagagcg tacccacata tgtacacacg cacacacaca cacacacctt  +RSINE2 445-477
 481 attttatcat aactataatc ctctggctcc atgactactt cataacaact ttaatcctcc
 541 tggctccaat actctcttac caccctctcc tcctggatcc taatactgga cacaaatatt
 601 taatccaaat ccaattttgt gtttgtcaat aatcttcagt gtctccctct cctcacaggg
 661 acaggagatt acatttcctt gtggtttatg acccctctg tcccagtact ggtagcaatt
 721 tatatacatg gcatgggaat atgttatatg tgtatggggc tgtgtgcaca tatatatgag  -RSINE2 742-881
 781 ggcacatgtc atcctatgtg tggaggccag aggtcaatgt cacgtcttcc ccaatcactg
 841 tccagtggtc cctggattcc aaactcaggt cctcatgctt gggaaccaac ccagcgccct
 901 gattcctaac ccccttctct ttcaaacaag gtctcattat gttgcctggg tcagccttga
 961 acttgagtct cctgcctcga cttctctgat tgcaggtttt cacctgtcca agtcagcagg
1021 catcttgaac aagagcatca tttcccttaa gctgcttcag gcggtgttca tgggggctct
1081 taaatgtagt gtacttttcc tttggacaca gtaaaagcaa aacaagatta aacgagctgg  +B1 1136-1268
1141 acacggtggt acatgtcaat aatctgagaa ctcaggaagc tgaggcagaa gcgttgcttt
1201 gagtgtgagg tcagcttgaa ctacttagag ggtagaacag gccaacttgg gctaagagac  -MER17A 1275-1342
1261 tctgtctcca aacataaaga aaaaagggag ggagagaaaa agaaaggaaa ggaaagaagg
1321 gagagagtag gacatctaag agcctgaatg ttggtggtgt tgaccagcca gactgaattt  5' UTR of cyt c; intron 1 out-splice
1381 aaaatgggta atgttgaaaa aggcaaggag atttttgttc agaagtgtgc ccagtgtcac  ATG start of cyt c pseudo ORF
1441 actgtggaga agagaggcaa acatgagact ggaccaaacc tccgtggtct gtttgggtgg
1501 aagacaggcc aggctgctga attctcttac acagacacca acaagaacaa aggcatcacc  intron 2 of cyt c out-splice
1561 tgaggagagg ataccttgat ggagtatttg gagaatccca aaaagtacat ctctggaaca
1621 aaaatgatct tcactggaat taagaaggga gaaagggcgg acctaatagc ttatctttaa
1681 aaggctgcta atgagtaatt ctactgtctt atttattaca aaacaaatgt cgtatggctt
1741 ttaatgtata ccataattta attcatacac caaattcaga tcatgaatgg ctaacaatgt  end of cyt c ORF
1801 ttttgttgga cagccttgat ttaagtaaaa ctgacctgtc ataggtggac atgatctctt
1861 ttaaagtaac acttccaatt gaatacatgc tatcactgct ctccctttct caagataaga
1921 ctggacttaa ttagtaatgt tttactttcc ataaatatgt ggctgtcacc tcaaacctat  350 bp of 3' UTR homology
1981 tgagtgtttt tatacttaga tctgtataac tgggtctatg aatatgttta aacactggga
2041 aaattctatc gctgtctcag aaacaagaag actcacccgt gcttcagtgt gtgttcactg
2101 gcctctgaca ggcaatgcta aacaccagga aggaactatt cttgacattg ctgttttaat
2161 tagaattccc tggggctggg gatttagctc agtggtagag cgcttaccta gaaagtgcaa  interrupting +RDRE1_RN 2174-2266
2221 ggccctgggt tcggtcccca gctccgaaaa aaaagaacca aaaaaaaaaa aagaattccc
2281 tacatcagtg atgctgcctt tCACCACtga aaggcattta cggtggttta tgtatgatat  center of stem loop before poly A signal
2341 caAATAAAga gtatttaaca cttctttata gttgaaaaag aaaggaaaga agggaatcaa  poly A signal; end 1100 mRNA; new poly A
2401 ccgagaatga taaaccaaca ttcaatggcc aatatacttt ctaagcctct aattctttta
2461 tagtttatgg ggaaatgtca aaaatcttcc tctttaccaa tttcttgtta ccaaagttcc  452 bp residual prion promoter
2521 acgatggctt tttctttccg ttaggtaacc tttcattttc tcgactaccc attatgtaac
2581 gggagcgctg ggttctggat cagtcttcca ttaaagatga cttttatagt ctgtgagcgt  conserved motifs
2641 cgtcacagag tgctgacact ggggtgggga ggggagtacg gggggagggg gttaaacaga  unknown insert in rat with flanking gggg
2701 taacaagcat ttaagccagt acggagcggt gactcatccc accgcgagaa gccattggtg  SP-1, AP-1, AP-2 CCATT sites
2761 agcatcacgc tccgcccctc gccccgccca gccccccggcc tgtcgggtcc ctcaccacgc
2821 cccgctcccc cgcgttgtca gagcagcaga cggagtctga gcgtcgcgtc ggtggcaggt  exon 1 of rat prion; start of intron 2
```

Genes são transcritos e traduzidos dando origem a proteínas

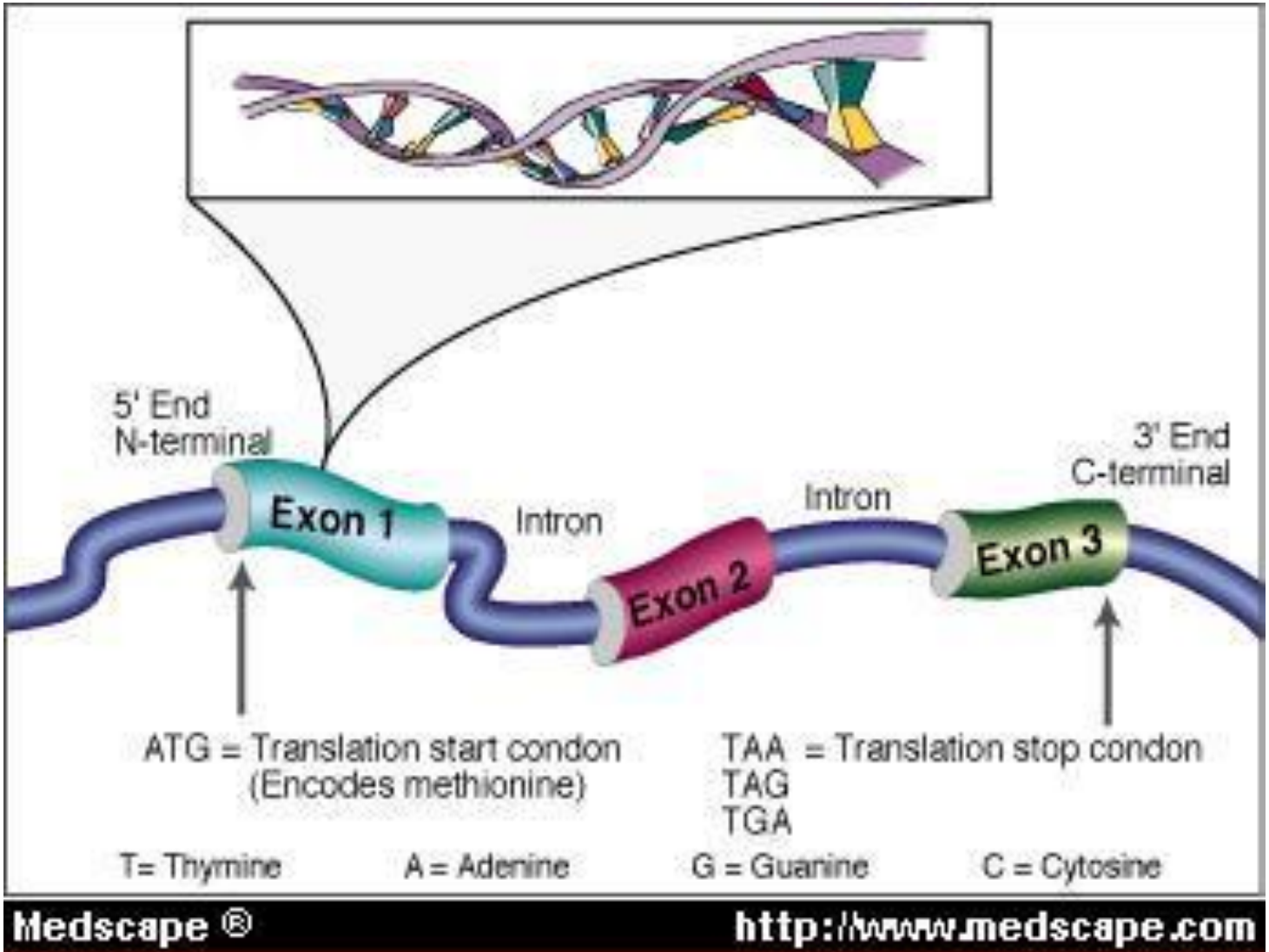


Mas se fosse muito simples não seria tão interessante

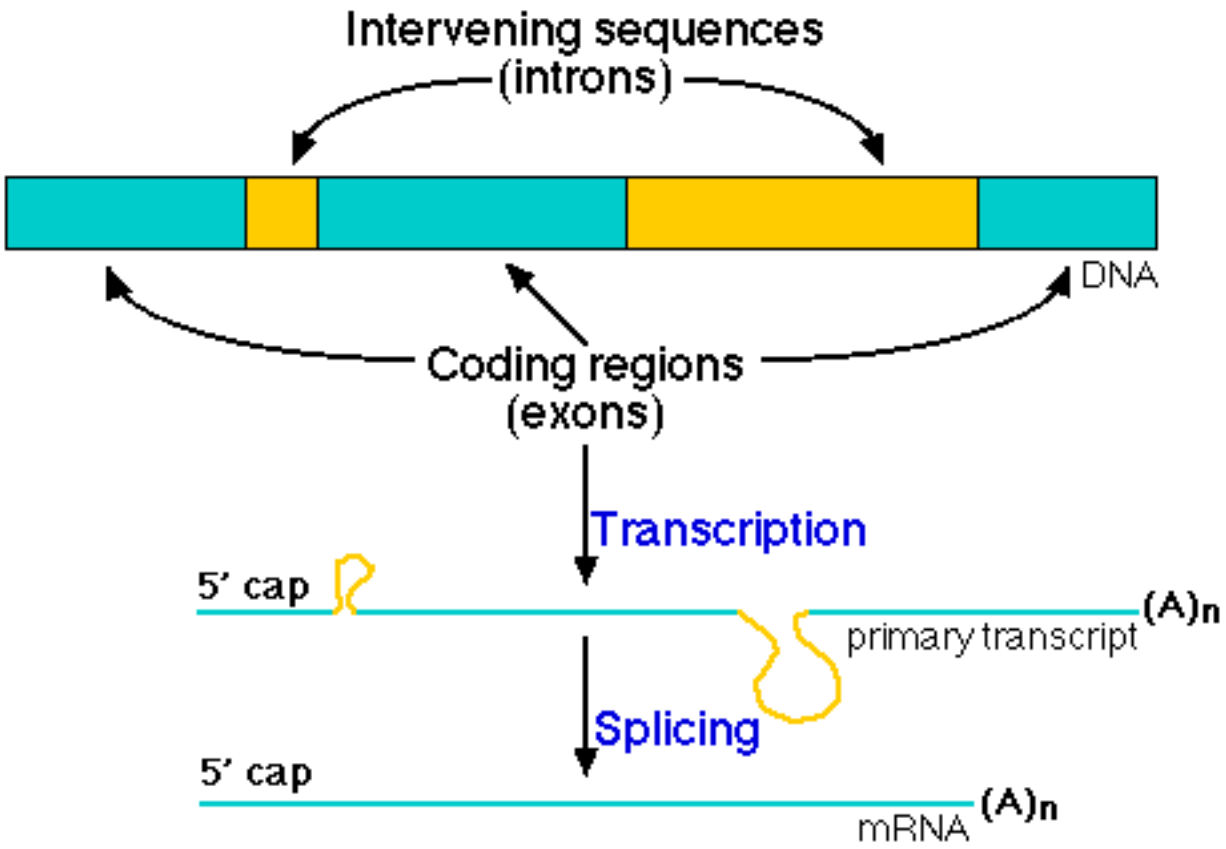
Foram encontrados apenas cerca de 25mil genes, ao invés dos 60 a 100 mil genes estimados a princípio.



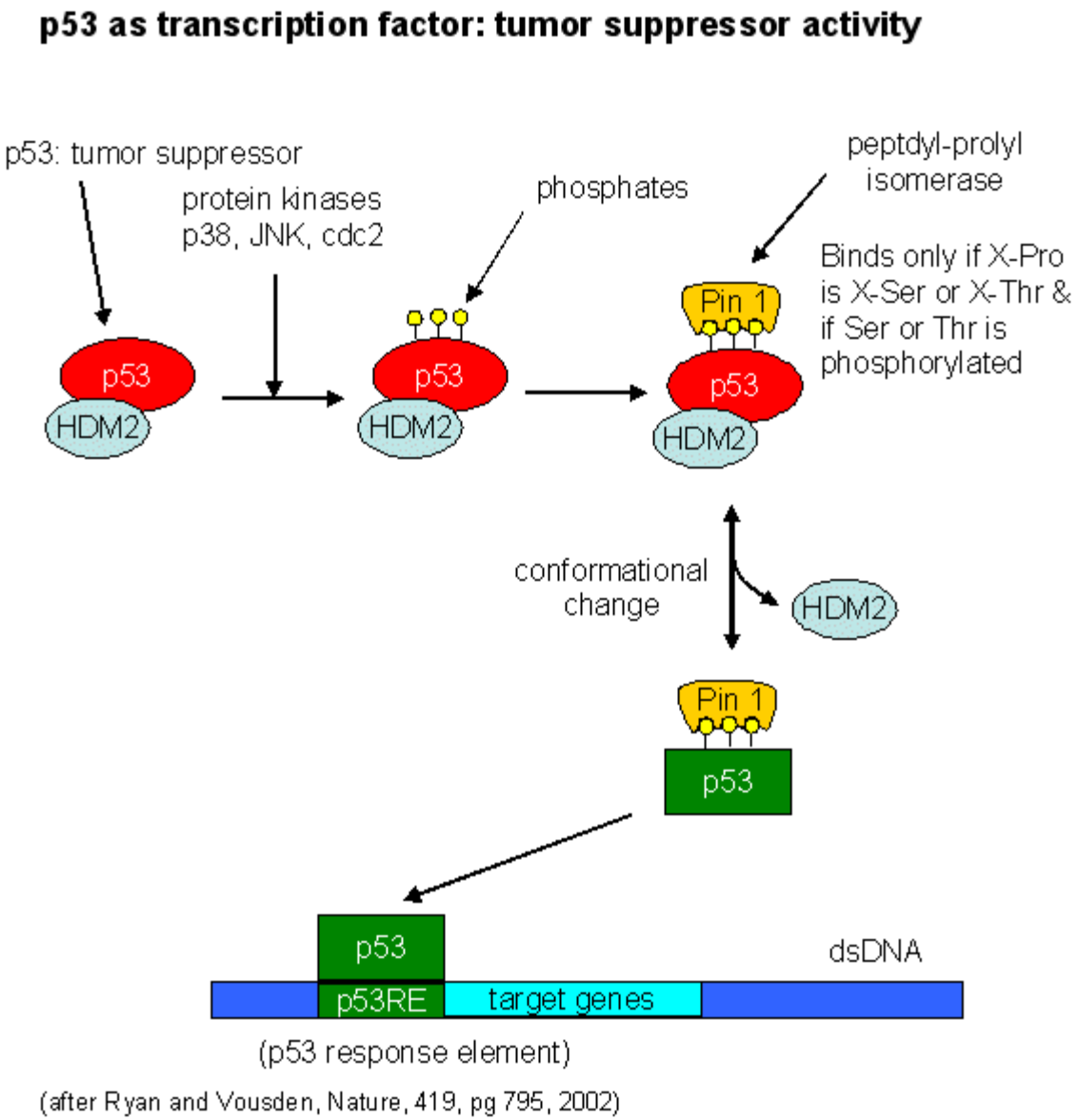
Os genes dos eucariotos são compostos por exons e introns …



Os introns sempre desaparecem por ocasião do splicing.
Os exons podem permanecer ou não.

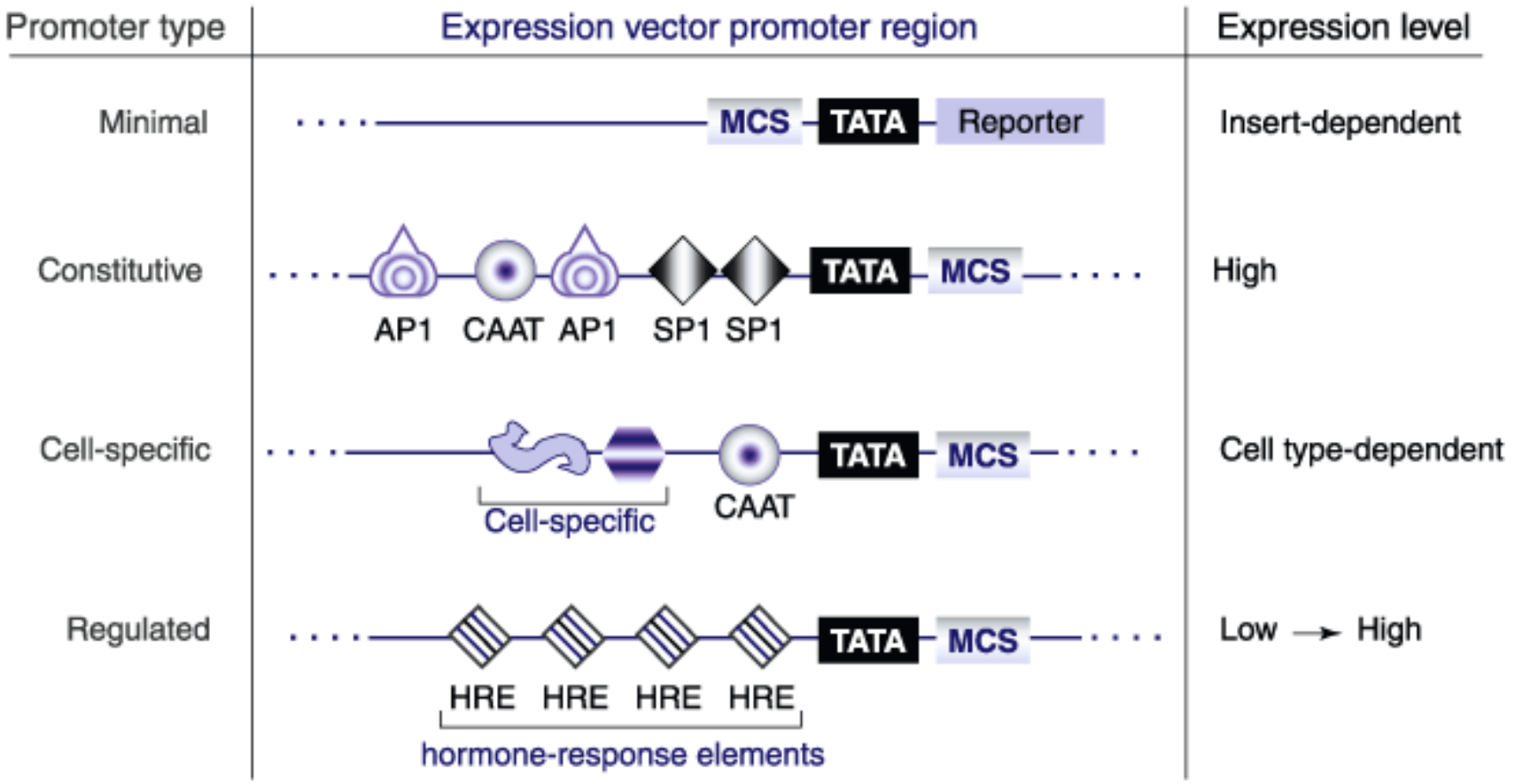


Os genes têm a sua expressão controlada por Fatores de Transcrição



O foco da atenção passou então para as proteínas.

Os níveis de expressão de um gene variam com os promotores que se ligam à região reguladora, imediatamente antes do gene



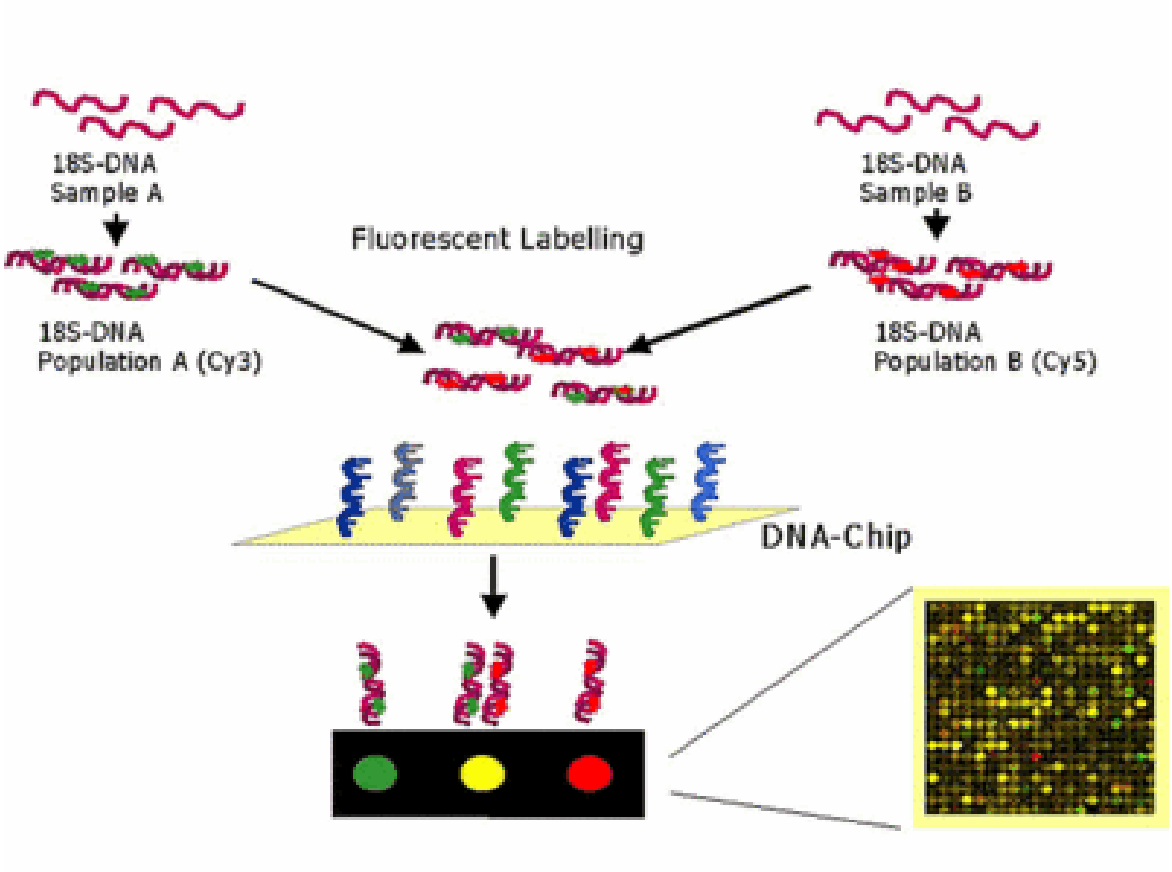
O que é Proteômica?

Proteômica é o estudo das proteínas, com suas estruturas e funções.

Link para Expasy:
http://ca.expasy.org/cgi-bin/prosite/PSView.cgi?ac=

Link para PDB:
http://www.rcsb.org/pdb/home/home.do

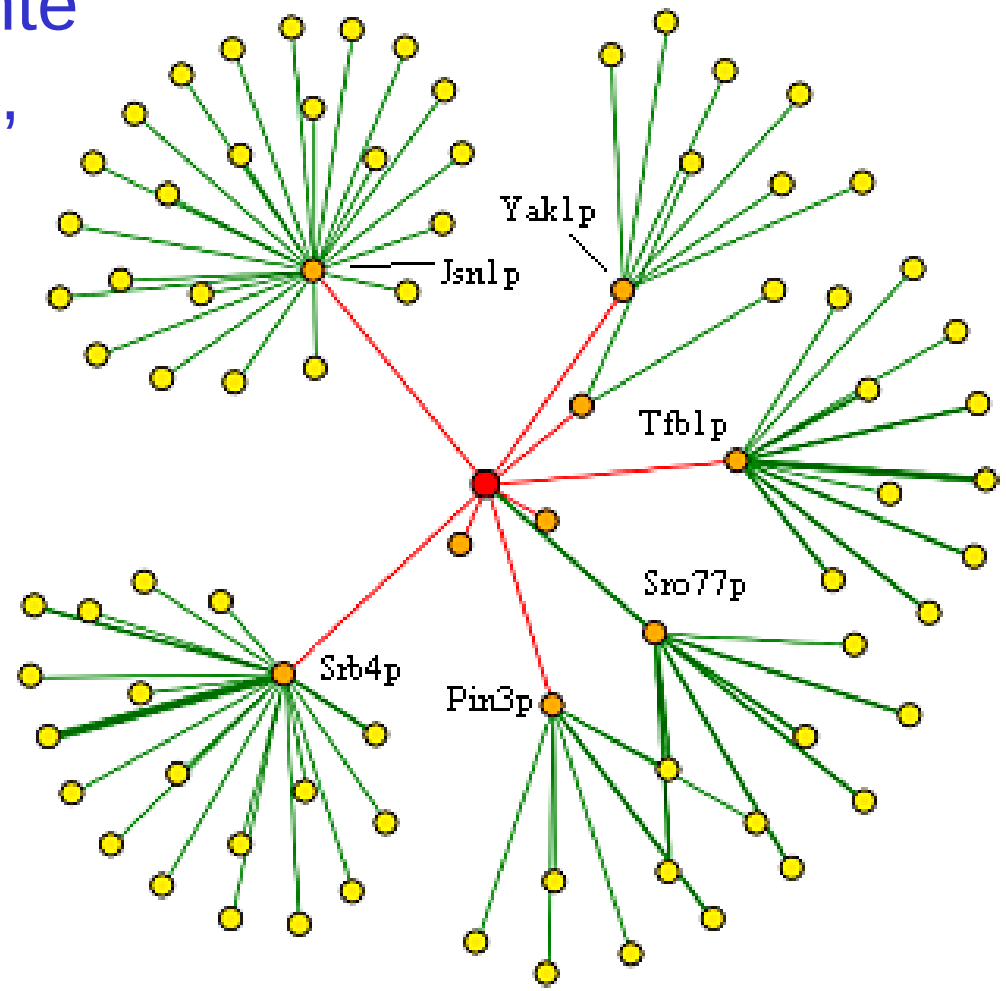
O nível de expressão dos genes em diferentes condições ou em intervalos de tempo pode ser medido



Em setembro de 2006, o GEO (Gene Expression Omnibus) do NCBI continha mais de 3.2 bilhões de medidas, tomadas sobre mais de 200 organismos.

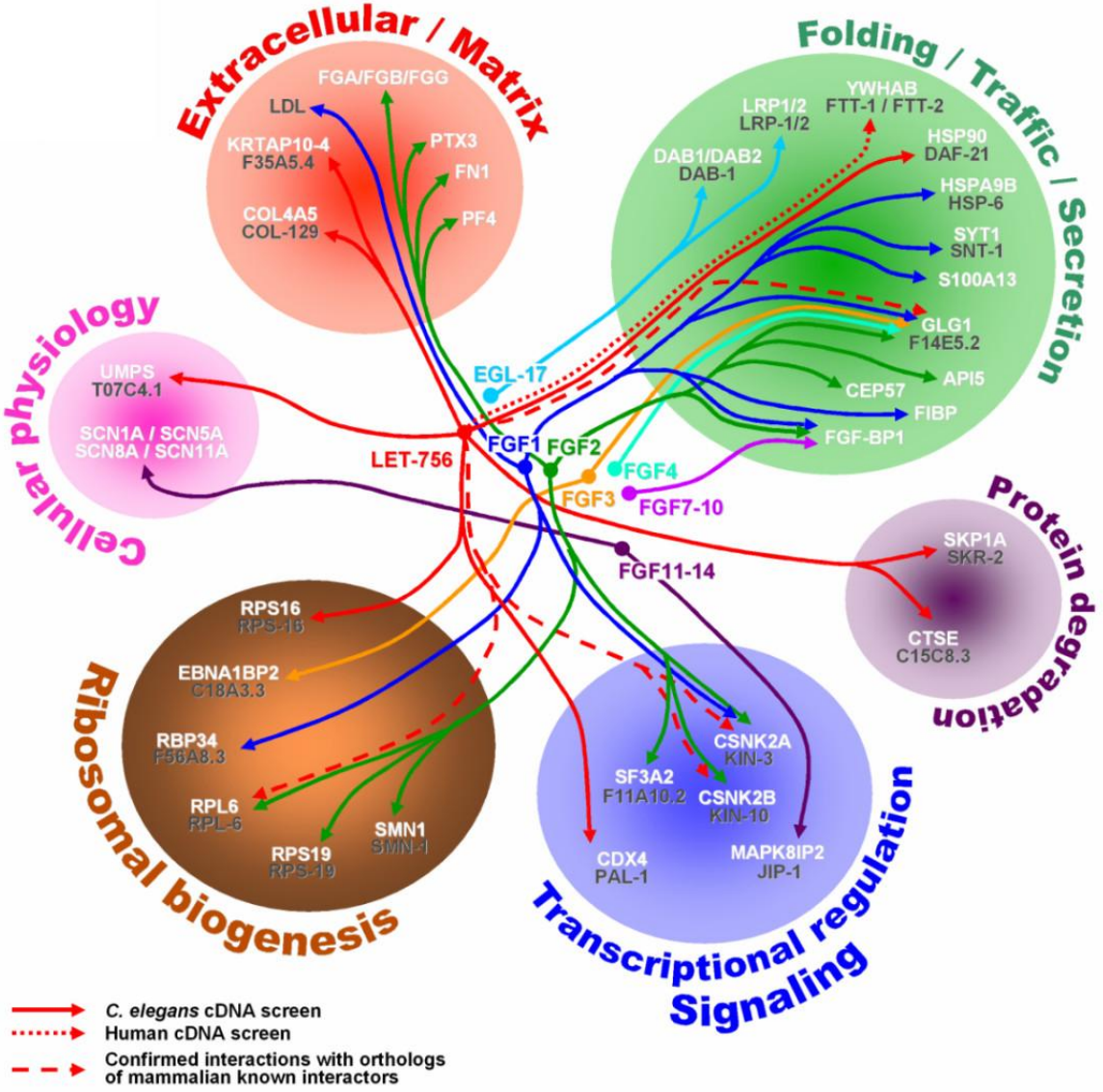
Interações entre Proteínas

Proteínas geralmente atuam em conjunto, e se organizam em redes do tipo *small world*, com muitos nós de grau baixo e poucos nós de grau alto (*hubs*)

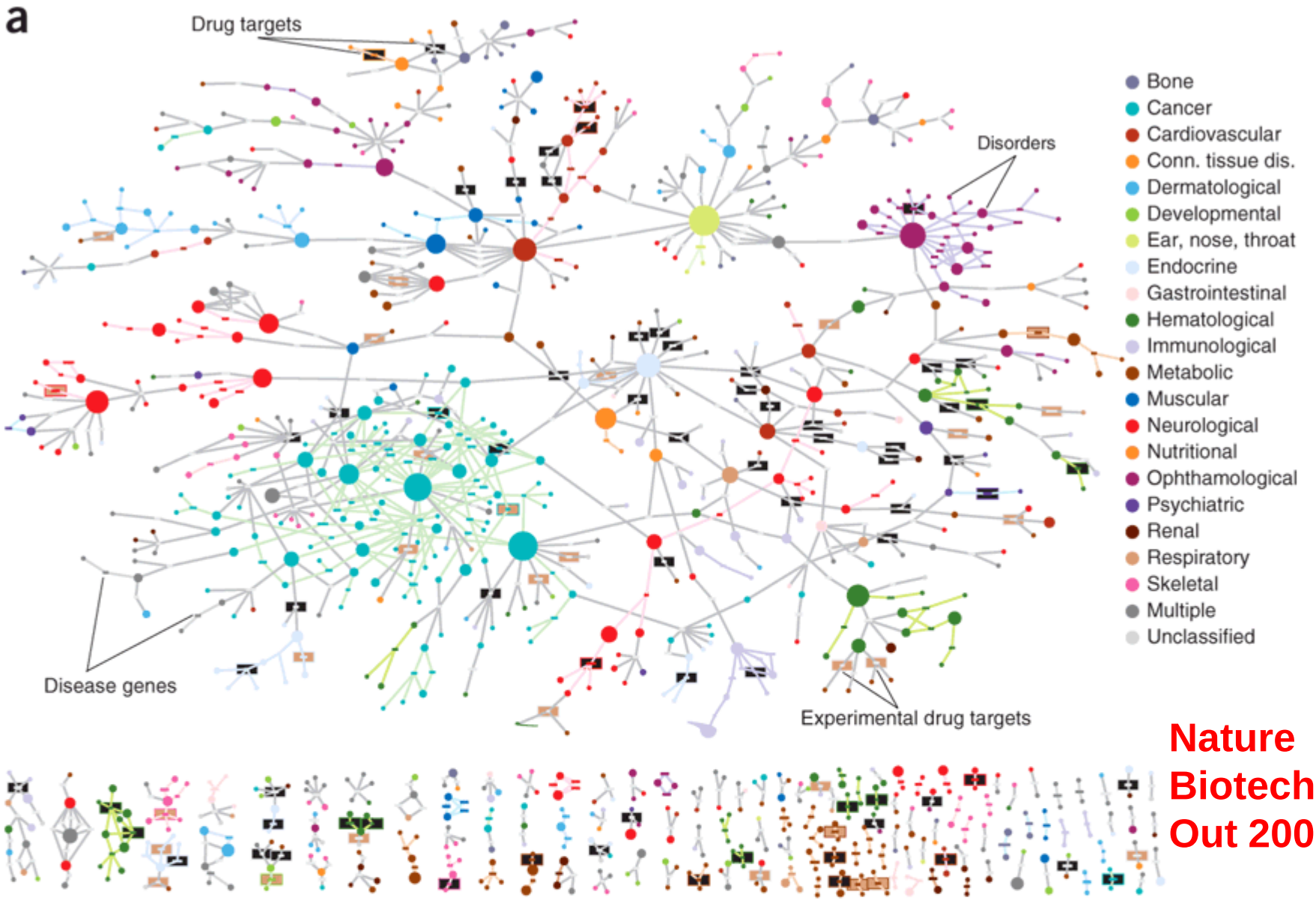


Interações entre Proteínas

Proteínas relacionadas funcionalmente encontram-se a uma distância muito pequena, e em geral são vizinhas nos mapas de interação.



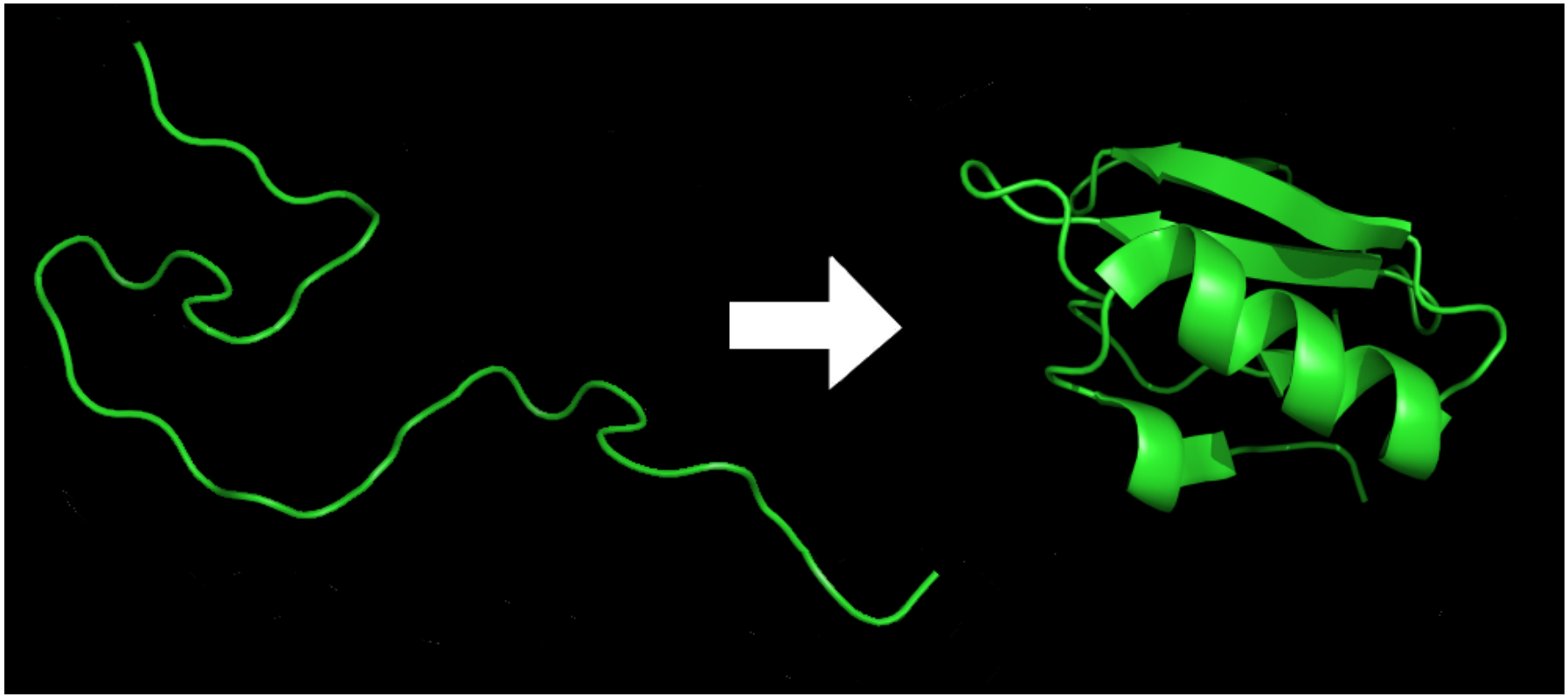
Doenças Humanas e Alvos de Drogas



Nature Biotech, Out 2007

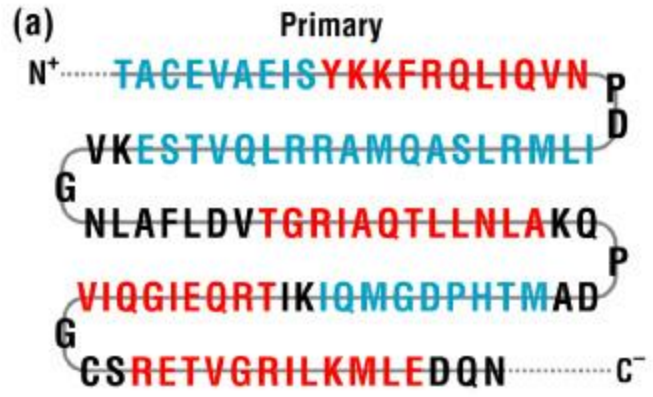
Dobramento de Proteínas

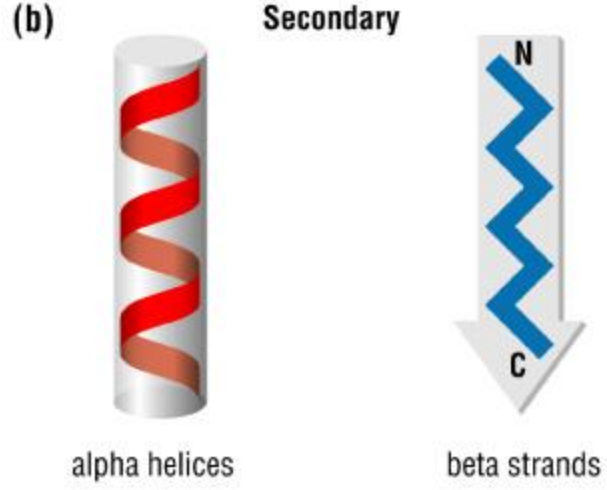
Um problema bem Difícil

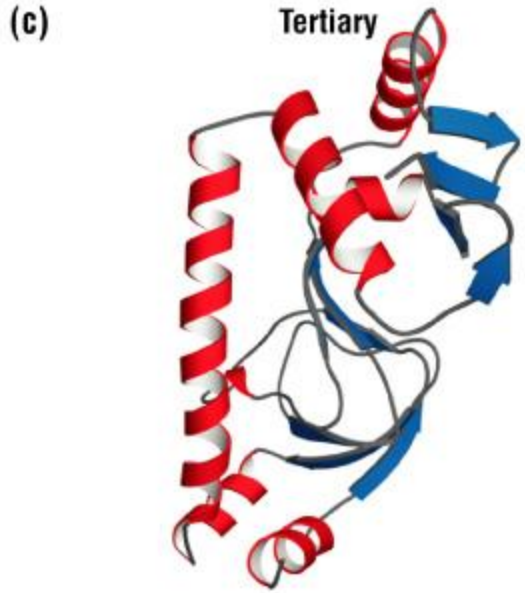


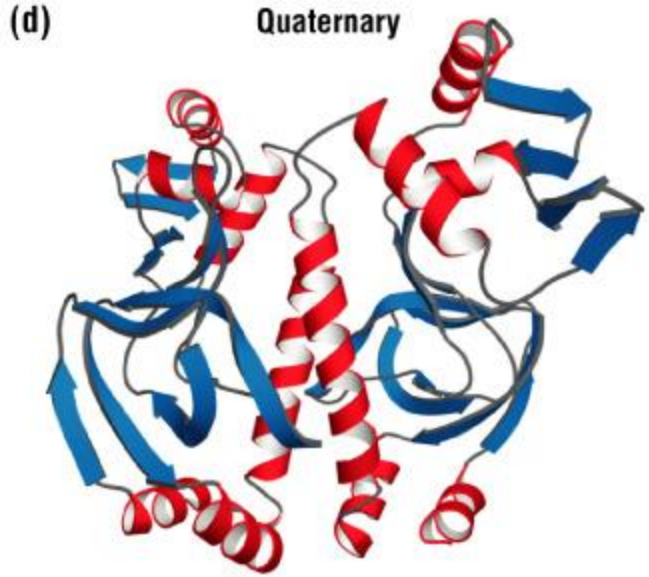
Predição de Estrutura de Proteínas

From **Protein Structure and Function**
by Gregory A Petsko and Dagmar Ringe

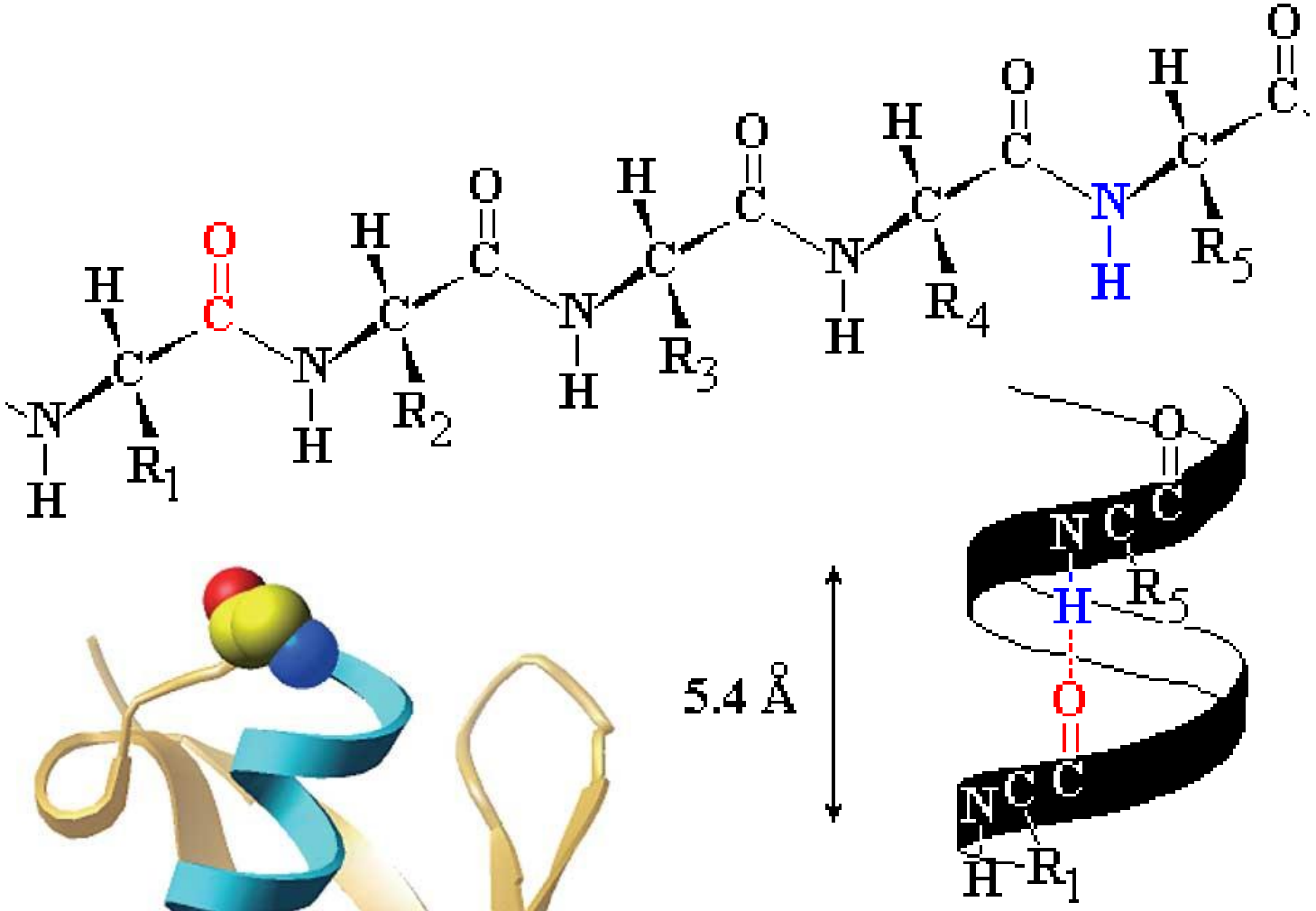






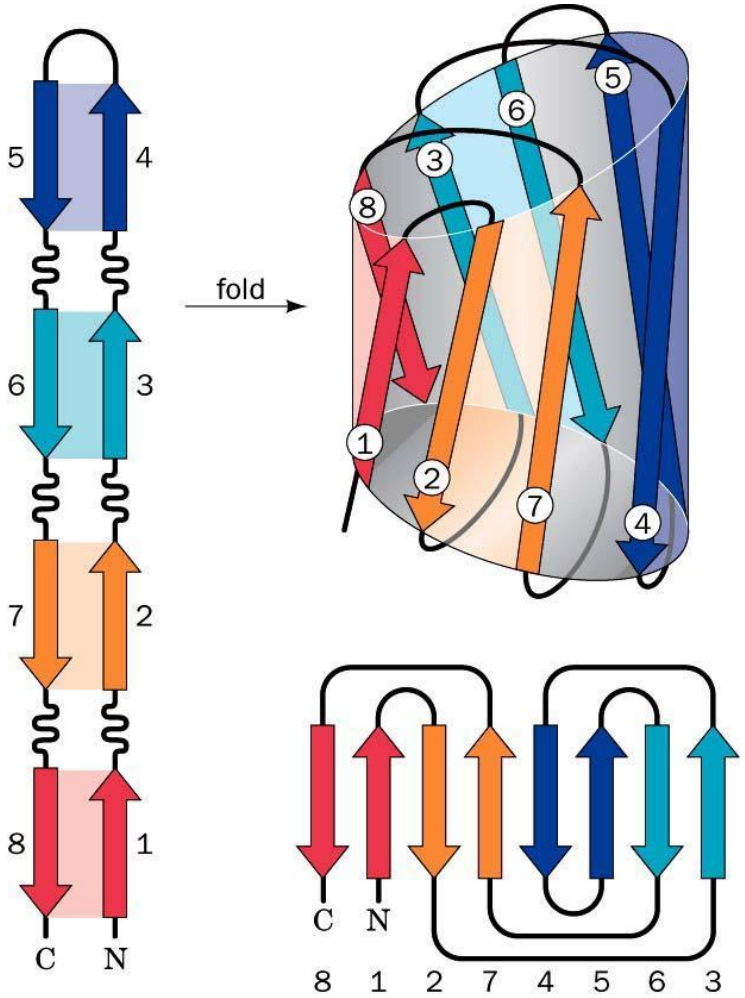


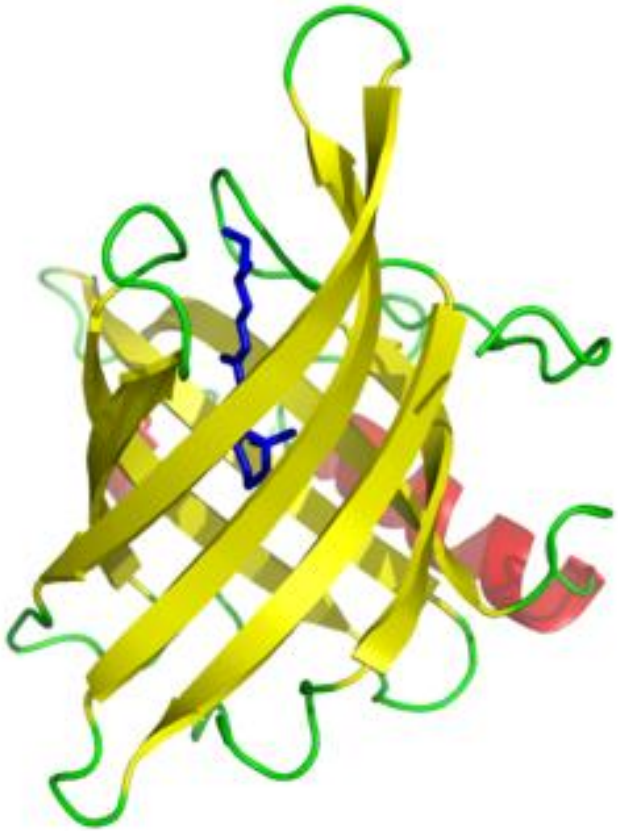
Dobramento de Proteínas



Alpha Helix

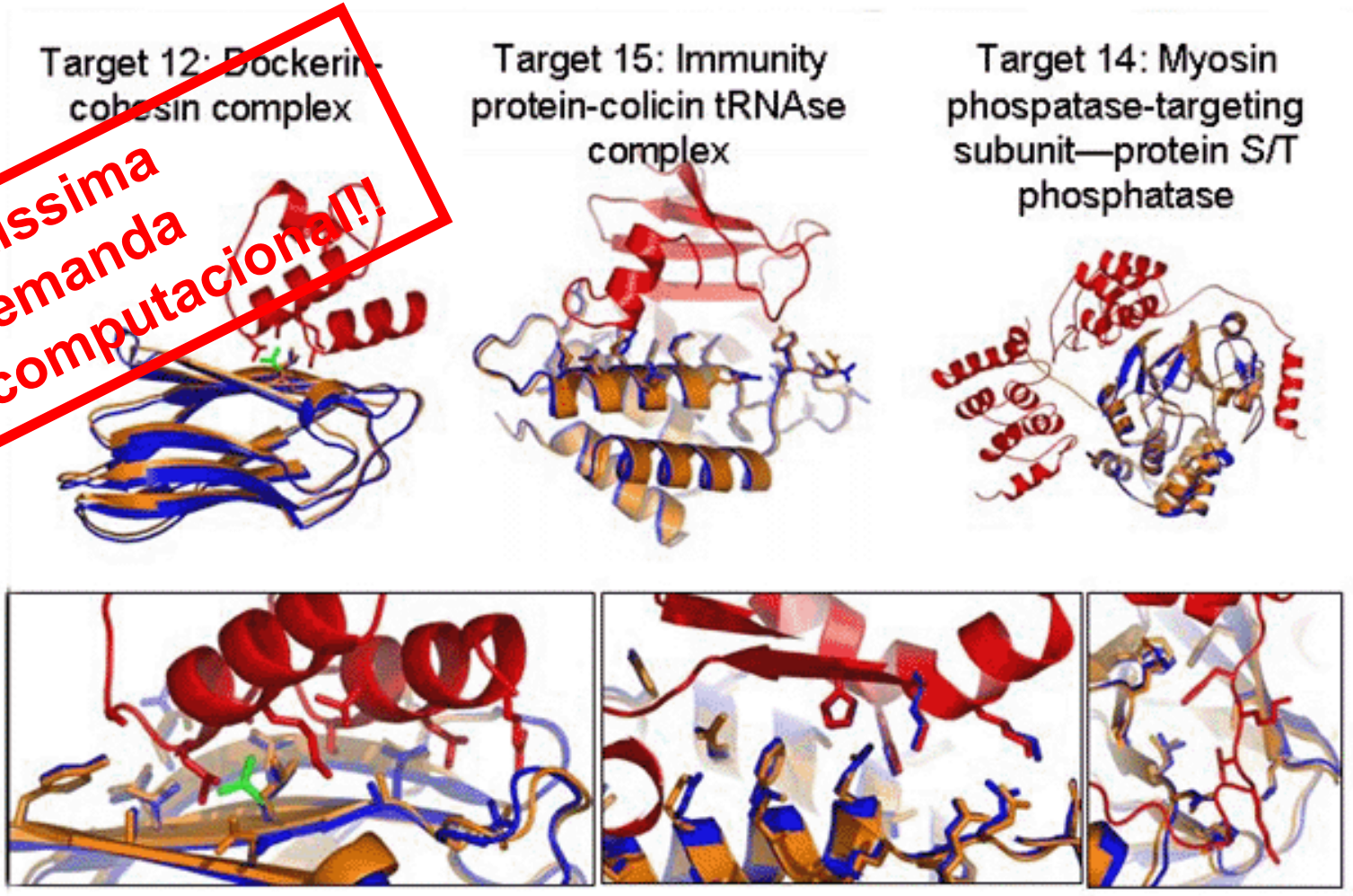
Dobramento de Proteínas
Beta Barrel





Docking e o Projeto de Drogas

Detalhes como orientação e ângulo de ligação de todos os resíduos do sítio ativo são essenciais.



The Blue Gene Project

Em dezembro de 1999, IBM anunciou um projeto orçado em $100 milhões de dólares em 5 anos.

Objetivo: Construir um computador massivamente paralelo para ser aplicado no estudo de fenômenos biomoleculares, como protein folding.

IBM Research – Blue Gene supercomputadores que operam da ordem de 478 TFlops (continuado) e 596 TFlops no pique!

Top 500 parallel Computers, Nov/2007: http://www.top500.org/

Nosso Trabalho Nesta Área

Nos últimos anos temos trabalhado em:

- Predição de Estruturas de Proteínas
- Reconstrução de Redes de Genes (em geral e relacionadas a doenças)
- Interação entre Proteínas e Genes
- SNPs (single nucleotide polimorphism) e Haplotyping

Novo!

Predição de Estruturas de Proteínas

IEEE

BIBE 2003

Combining Few Neural Networks for Effective Secondary Structure Prediction

K. Guimarães, J. Melo e G. Cavalcanti

Bethesda, USA, Março 2003

Protein Secondary Structure Prediction: Efficient Neural Network and Feature Extraction Approaches

J. Melo, G. Cavalcanti e K.Guimarães

IEE Electronics Letters, 2004

Reconstrução de Redes de Genes

IJCNN 2005

A Simpler Bayesian Network Model for Genetic Regulatory Network Inference

Gustavo Bastos and Katia S. Guimarães

Montreal, CA, Agosto 2005

Analyzing the Effect of Prior Knowledge in Genetic Regulatory Network Inference

Gustavo Bastos and Katia S. Guimarães

Dezembro 2005

Interação entre Proteínas e Genes

ALGORITHMS FOR
MOLECULAR BIOLOGY

Decomposition of overlapping protein complexes: A graph theoretical method for analyzing static and dynamic protein associations

E. Zotenko, K. S. Guimarães, R. Jothi, T. Przytycka

Abril 2006

Predicting domain-domain interactions using a parsimony approach

K. S. Guimarães, R. Jothi, E. Zotenko, T. Przytycka

Novembro 2006

Interação entre Proteínas e Genes

Interrogating domain-domain interactions with parsimony based approaches

K. S. Guimarães and T. Przytycka

Março 2008

SNPs e Haplotyping

… ataggtccCtatttcgcgcCgtatacacgggActata … → CCA
… ataggtccGtatttcgcgcCgtatacacgggTctata … → GCT
… ataggtccCtatttcgcgcCgtatacacgggTctata … → CCT

0.1% diferença de um indivíduo para outro.

80% das variações em SNPs

Ponto frequente de caracterização de doenças

Abordagens Combinatoriais e Estatísticos

SNPs e Haplotyping

Araújo, FRB ; GUIMARÃES, K. S. . A Case-Control Study of Non-parametric Approaches for Detecting SNP-SNP Interactions. In: XXX International Conference of the Chilean Computer Science Society, 2011, Curicó, Chile. Proc. of the XXX International Conference of the Chilean Computer Science Society, 2011.

SNPs e Haplotyping

Rosa, Rogério S. ; GUIMARÃES, K. S. . Insights on Haplotype Inference on Large Genotype Datasets. In: Brazilian Symposium on Bioinformatics (BSB) 2010, 2010, Búzios, RJ, BRAZIL. Lecture Notes in Bioinformatics. Berlin, Alemanha : Springer, 2010. v. 6268. p. 47-58.

Rosa, Rogério S. ; Santos, R.H.S. ; GUIMARÃES, K. S. . Accurate Prediction of Error in Haplotype Inference Methods through Neural Networks. In: IJCNN - Int. Joint Conference on Neural Nets, 2012, Brisbane. Proc. of the IJCNN 2012. Piscataway, NJ, USA : IEEE Publishing, 2012.

Micro Array Clustering

Monteiro, Carla C.R.R. ; GUIMARÃES, K. S. . Logistic Biclustering Models for Protein Network Inference. In: IEEE International Conference on Bioinformatics and Bioengineering, 2009, Taichung, Taiwan. BIBE 2009 Proceedings, 2009.