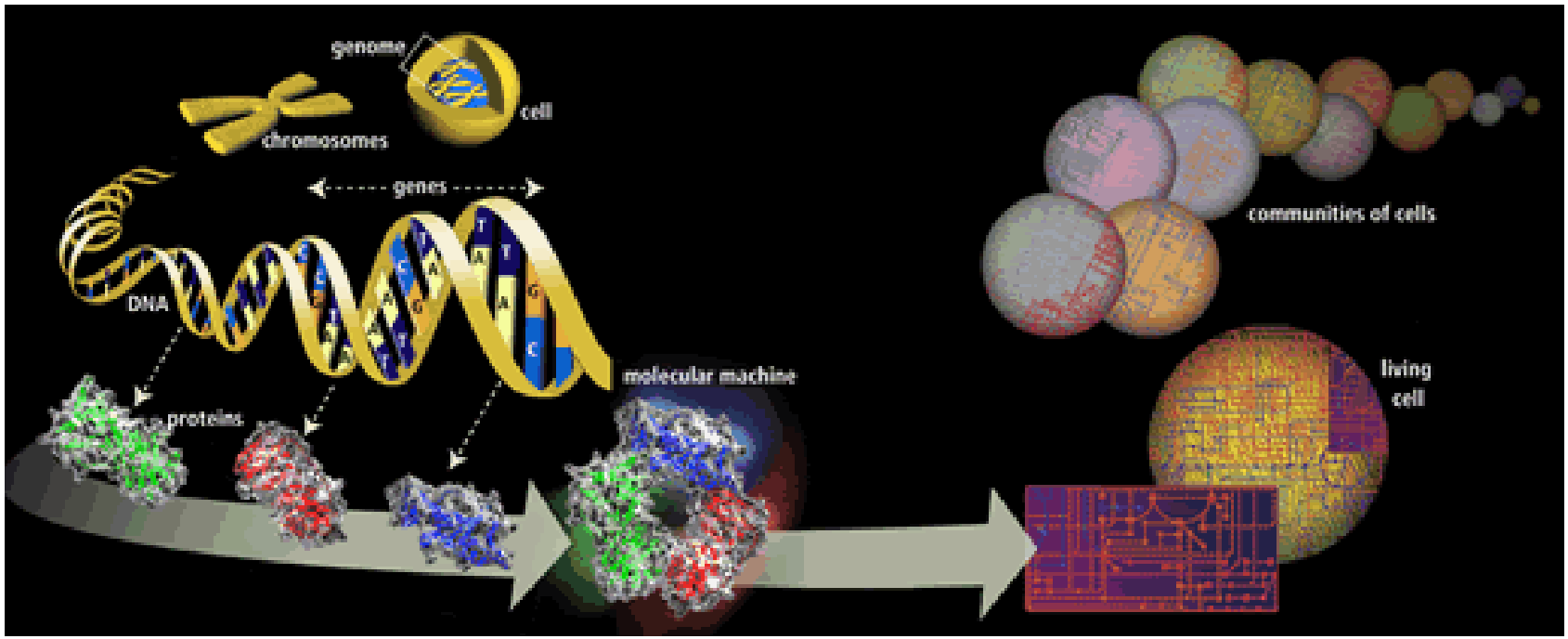
Miguel Rocha – Dep. Informática - Universidade do Minho

BIOINFORMÁTICA:
passado, presente e futuro !!



Bragança, 11 de Maio de 2006

Porquê a Bioinformática?

- Novas tecnologias experimentais da Biologia Molecular (e.g. projectos de sequenciação dos genomas) são capazes de criar **enormes quantidades de informação**.
- Estas apenas podem ser analisadas com recurso a ferramentas computacionais capazes de **extrair conhecimento útil dos dados**.

Porquê a Bioinformática?

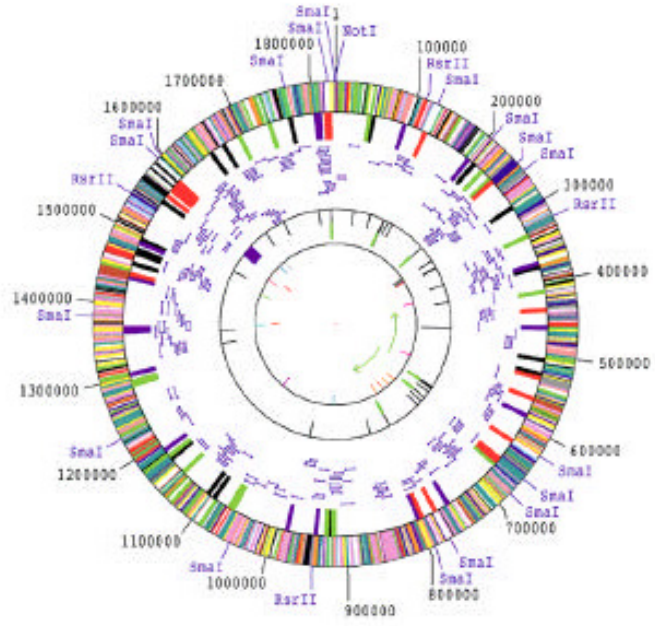
- Informação biológica mais **complexa** e **volumosa** coloca novos desafios aos campos das Ciências da Computação e da Optimização.
- Necessidade de **algoritmos** mais eficientes especializados na resolução destes problemas !!
- Sistemas biológicos são **sistemas complexos** – necessárias ferramentas poderosas na **modelação** e **simulação** destes sistemas (e.g. redes de regulação da expressão dos genes no interior de uma célula).

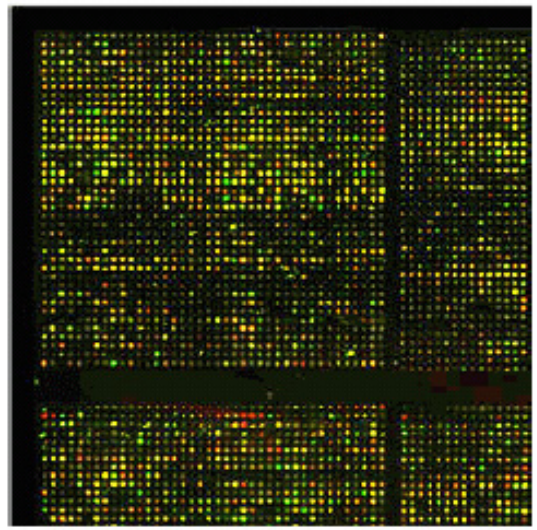
O que é a Bioinformática ?

- **Bioinformática** pode ser definida como o **armazenamento**, o **processamento**, a **análise**, a **previsão** e a **modelação** de dados biológicos com a ajuda das **ciências e tecnologias da computação**.
- Muitas outras definições (porventura igualmente válidas) podem ser encontradas na literatura, algumas mais abrangentes outras mais focadas.
- Termo idêntico, mas normalmente mais genérico: **Biologia Computacional**.

O que é a Bioinformática ?

Análise integrada da informação biológica em larga-escala





Estrutura **Genomas** **Expressão genética**

O que é a Bioinformática ?

Armazenamento, processamento, análise, modelação de informação biológica a diversos níveis

1 CÉLULA

1 TECIDO

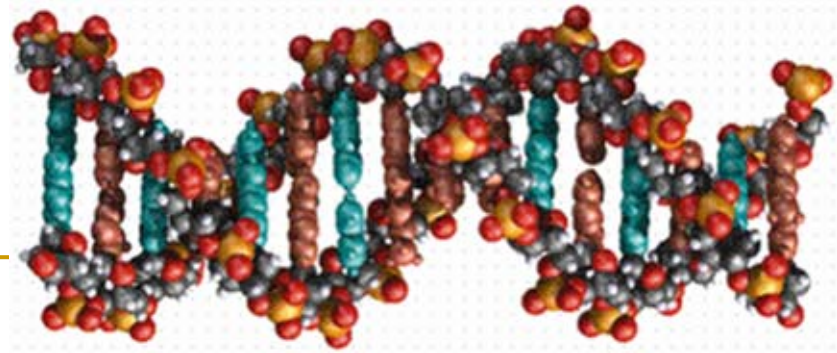
1 ORGANISMO

1 COMUNIDADE

1 ECOSISTEMA

Bio-Informação

- Desde a descoberta de que o DNA actua como um “livro” de instruções que comanda a vida, a Biologia tornou-se um mais uma **ciência da informação**.
- Muitos seres vivos foram já sequenciados podendo fazer-se comparação dos respectivos genomas.
- Estamos a aprender a ler o DNA !!



Moléculas da vida

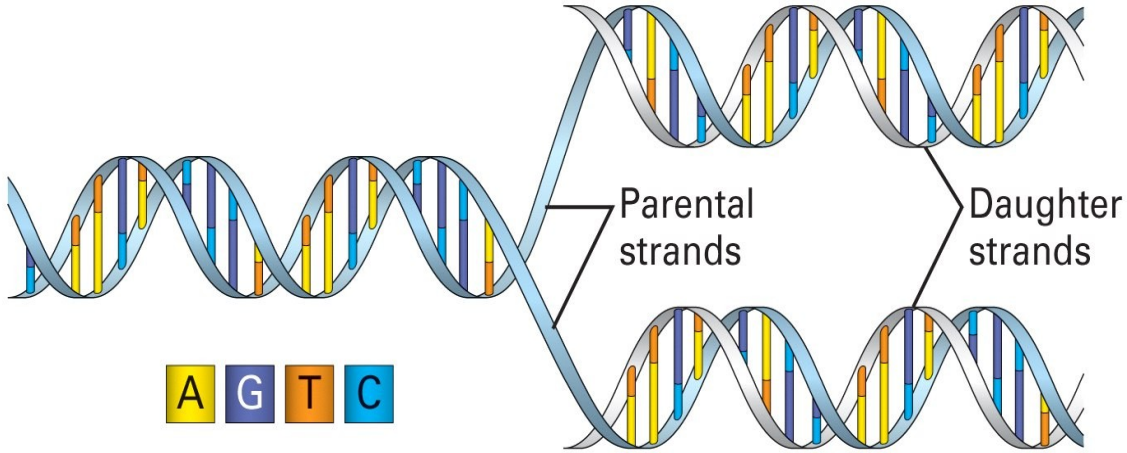
- **DNA**
  - Guardam informação sobre como a célula funciona (as instruções dos “programas” que regulam o funcionamento da célula)
- **RNA**
  - Transferem pequenos fragmentos de informação entre diversas partes da célula
  - Funcionam como modelos para a síntese de proteínas
- **Proteínas**
  - Enzimas que são usadas na sinalização intra- e extra-celular;
  - Regulam a actividade dos genes (regulação)
  - Constituem componentes estruturais do corpo (e.g. cabelo, pele, etc.)

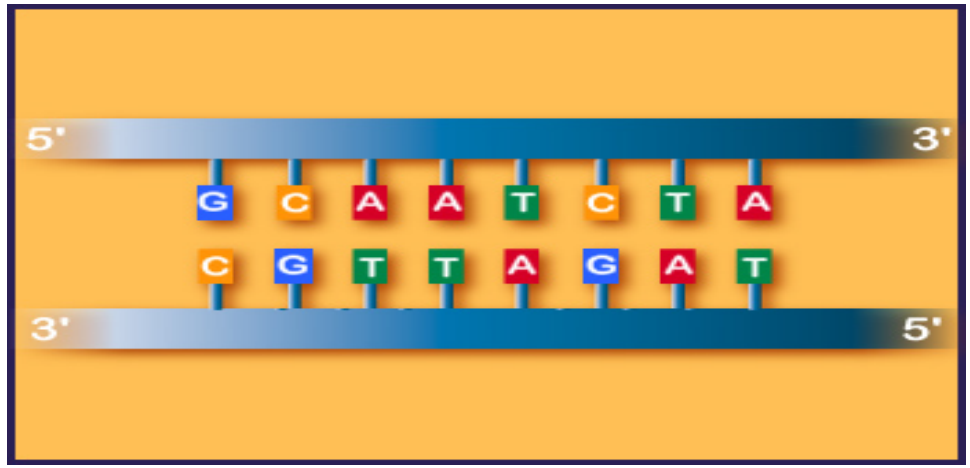
O livro da vida

- Tal como o texto humano o DNA, o RNA e as proteínas podem ser vistos como exemplos de **strings num alfabeto** de:
  - 4 letras – 4 nucleótidos do DNA ou RNA (A C G T/U)
  - 20 letras - aminoácidos nas proteínas.
- Muitos algoritmos em Bioinformática lidam com esta informação simplicada: biomoléculas são representadas e manipuladas como strings.

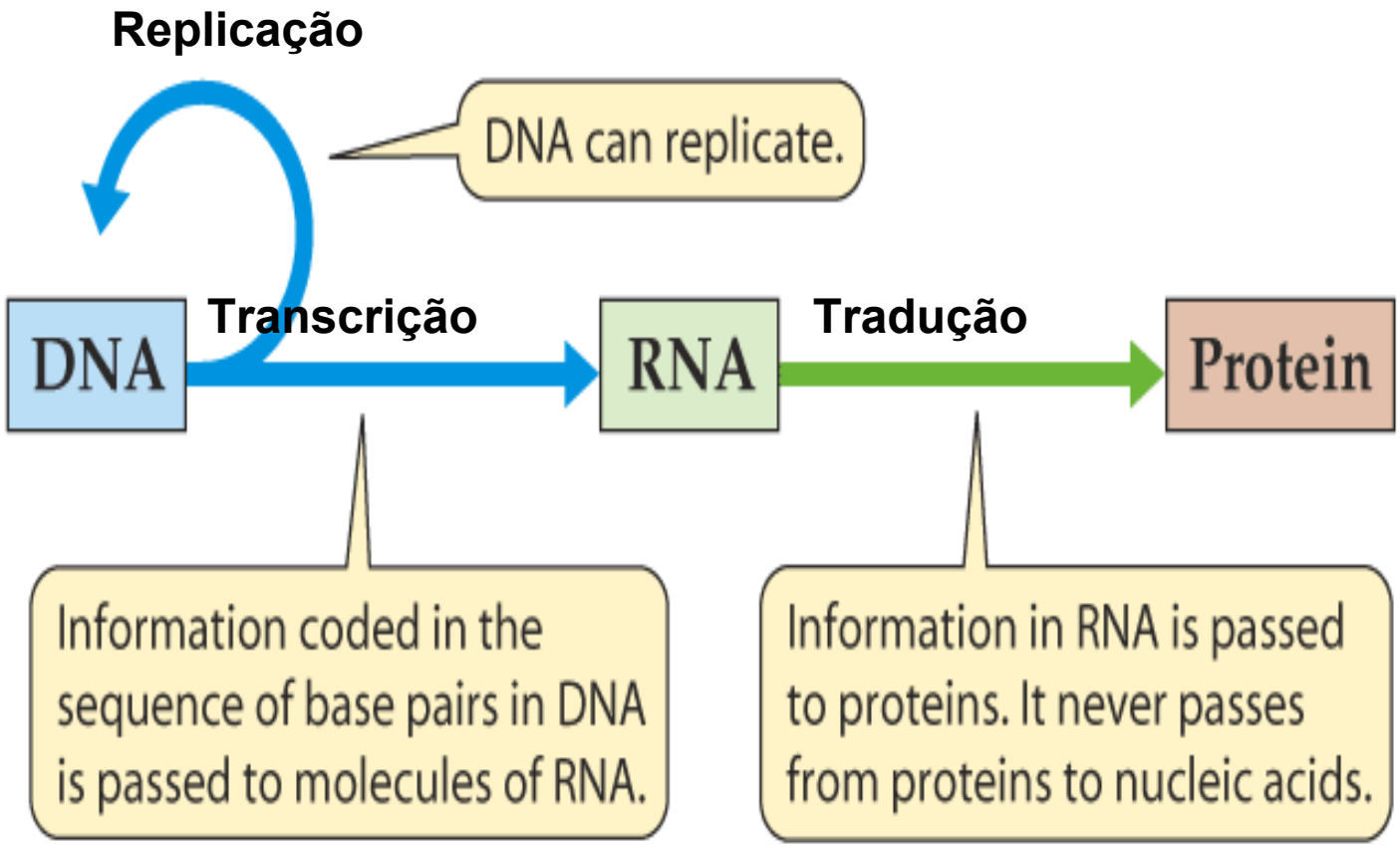
DNA: código universal da vida

- Guarda toda a informação necessária à vida **para todos os seres vivos**.
- **Adenina**, **Guanina**, **Timina**, **Citosina**, emparelham A-T e C-G em cadeias complementares

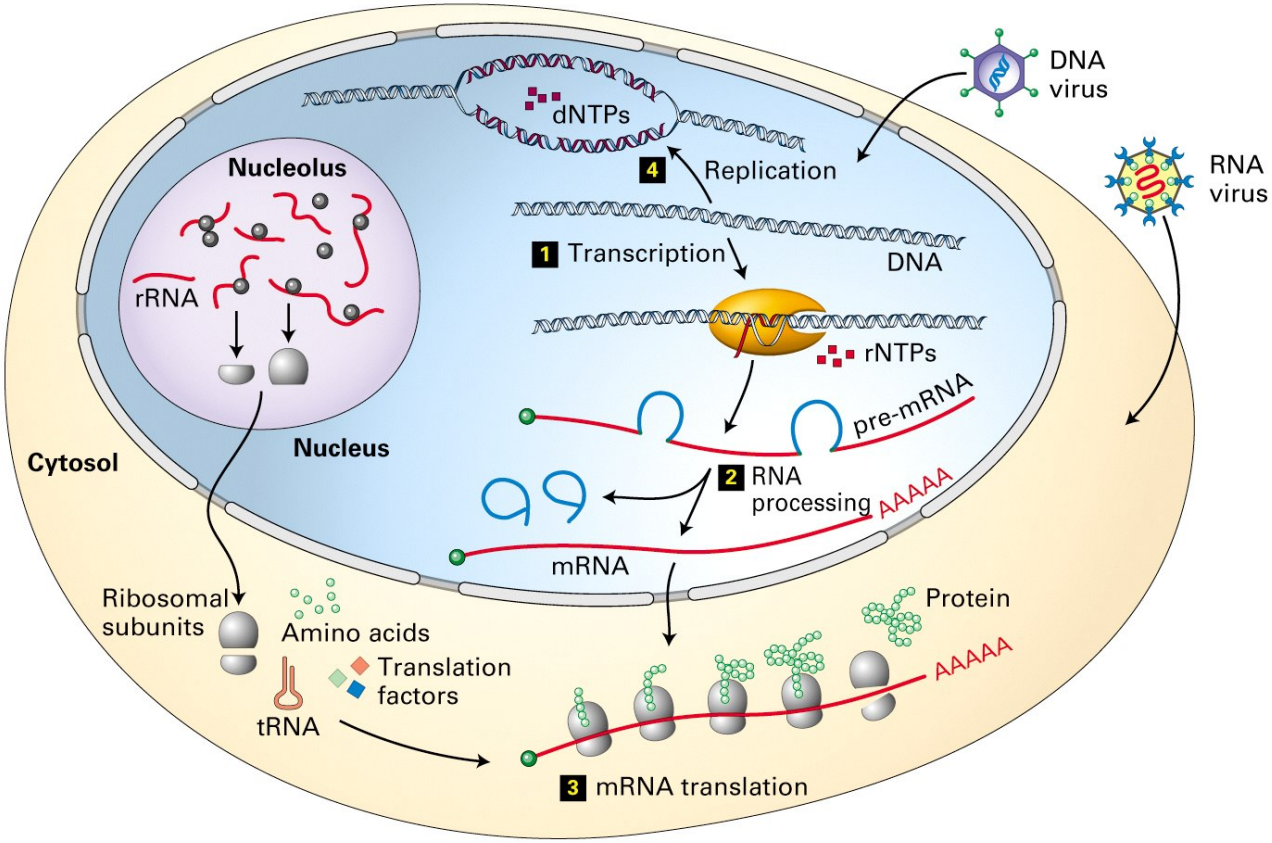




DNA, RNA e o fluxo de informação: o dogma central da biologia

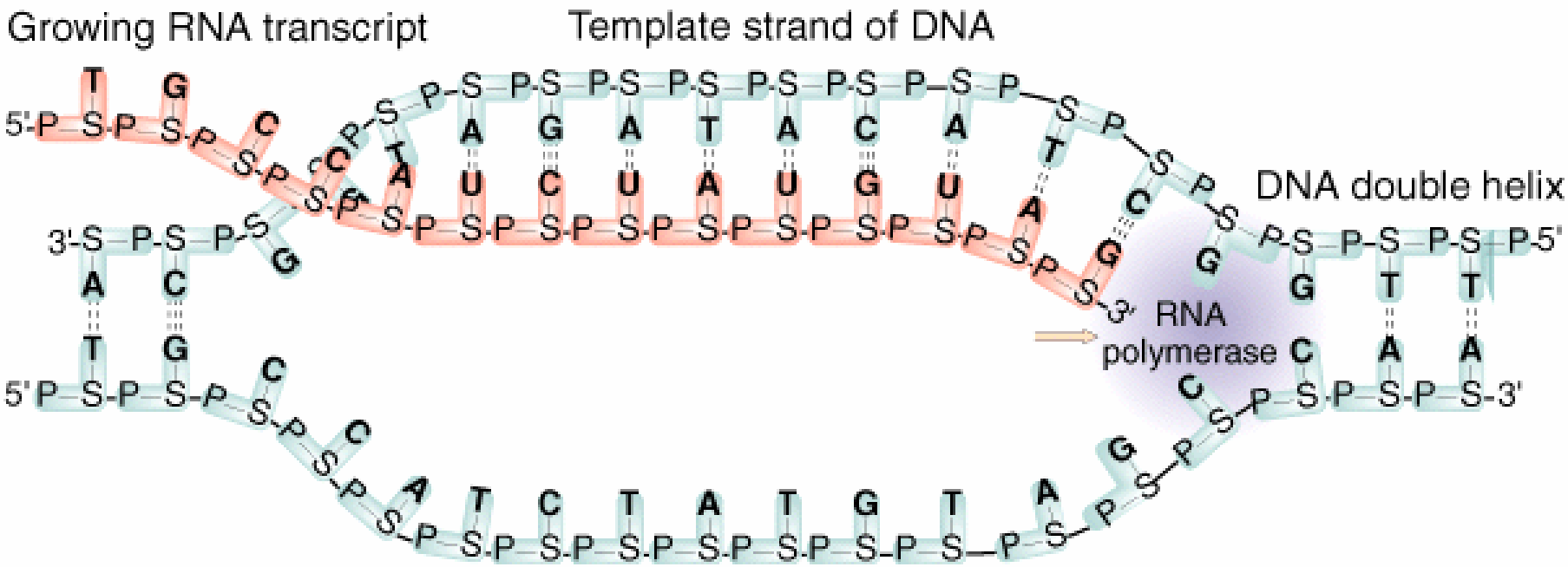


DNA para RNA para proteína



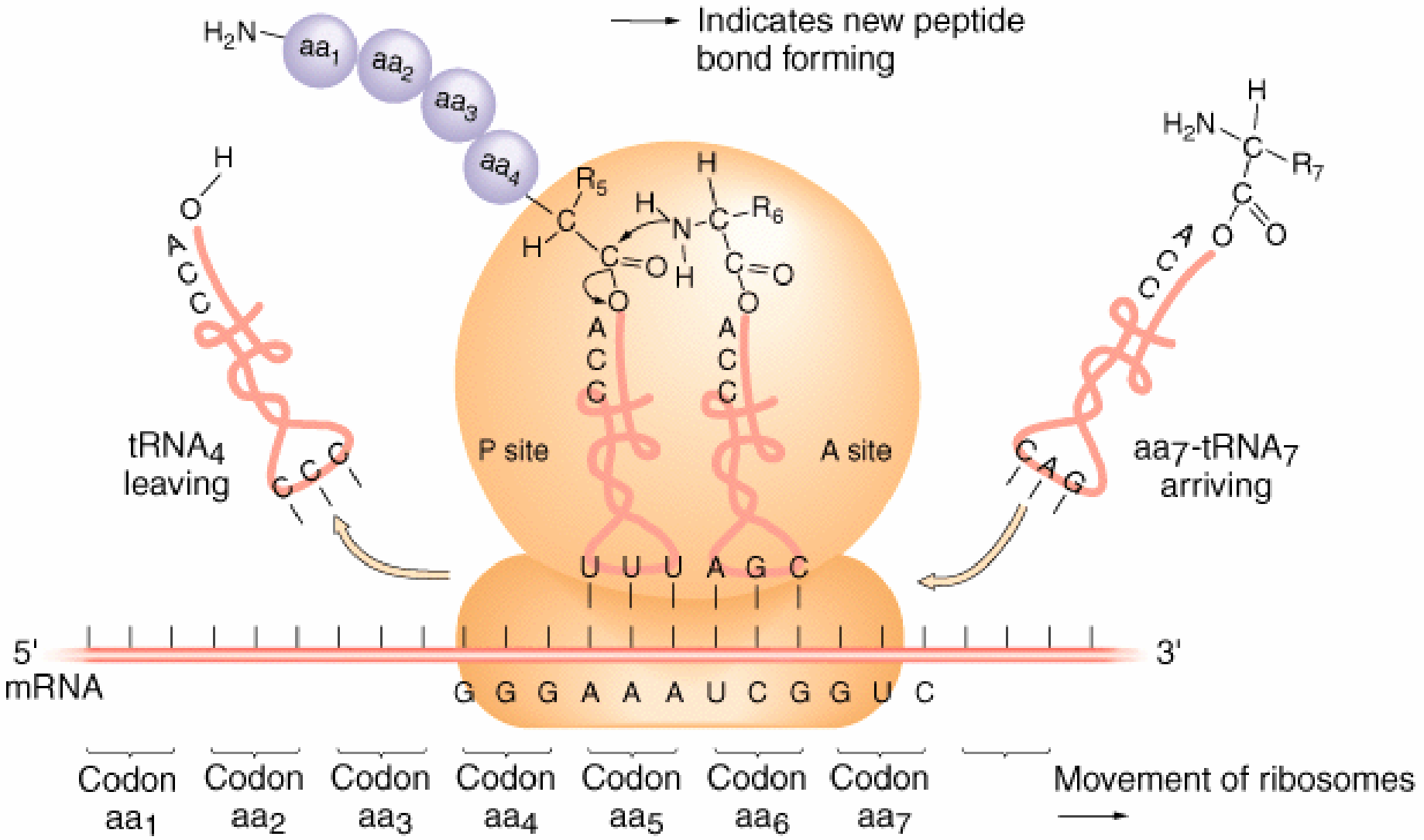
- Um gene é expresso em dois passos:
  1) **Transcrição**: síntese de RNA
  2) **Tradução**: síntese de proteína

Síntese de RNA: TRANSCRIÇÃO

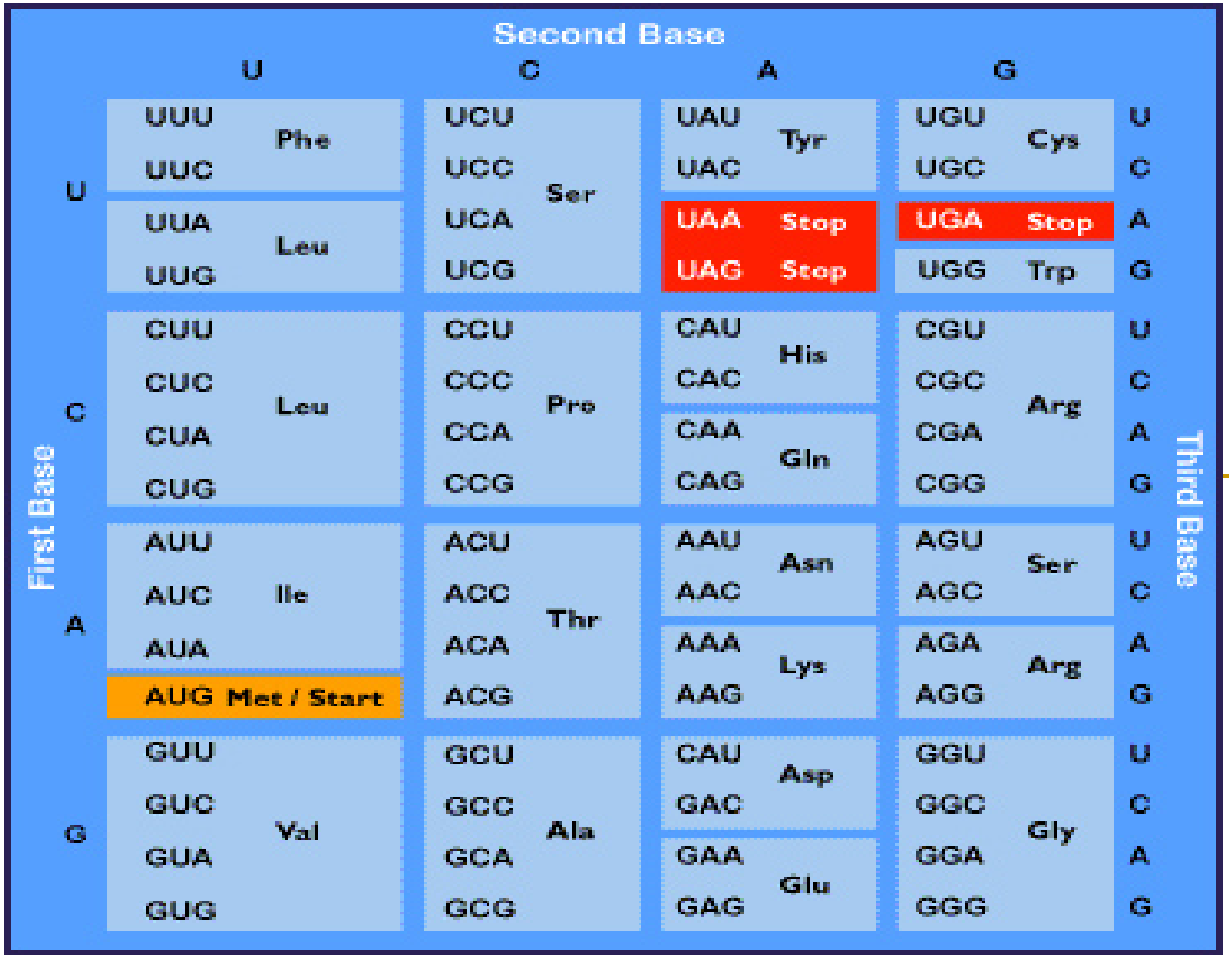


POLIMERASE DO RNA

Síntese proteica: TRADUÇÃO



Código genético



Papéis da Bioinformática

- Análise, armazenamento, processamento de **sequências** biológicas (DNA/RNA, proteínas) em larga escala;
- Organização e manuseamento de **bases de dados biológicas**;
- Extracção de conhecimento útil a partir de sequências biológicas (**Data Mining**): anotação de genomas, identificação de genes, previsão da estrutura e da função de biomoléculas, etc.

Papéis da Bioinformática

- Resolução de problemas complexos de **optimização**: alinhamentos de sequências e estruturas; identificação de padrões; inferência de árvores filogenéticas; agrupamento de genes pela sua expressão.
- **Modelação** e **simulação** de processos biológicos: modelação de processos metabólicos e regulatórios ao nível celular, de tecidos de células, de organismos !!

Bases de dados biológicas

- **Sequências de DNA, RNA**
  - GenBank (NCBI) http://www.ncbi.nlm.nih.gov/Genbank
  - EMBLBank (EBI) http://www.ebi.ac.uk/embl/
  - DDBJ (Japan) http:// www.ddbj.nig.ac.jp
- **Sequências de proteínas**
  - UniProt
  - Swiss Prot http:// www.expasy.org
- **Estruturas de proteínas**
  - PDB http:// www.rcsb.org/pdb

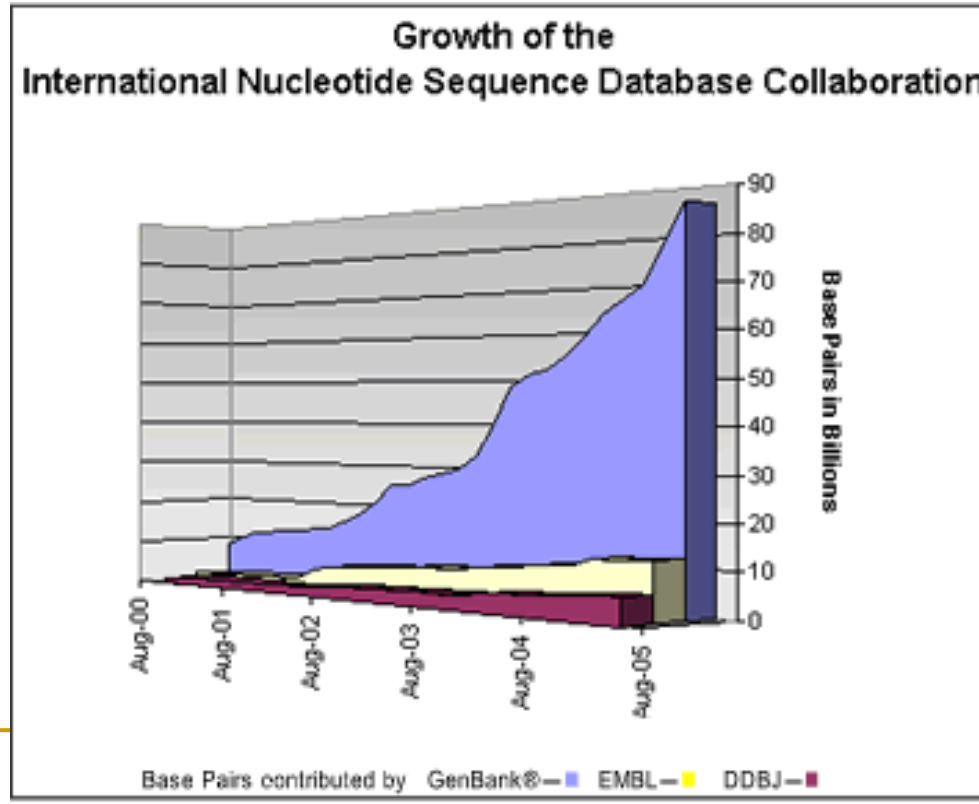
Bases de dados biológicas

- **Metabolismo** – reacções, vias metabólicas (e.g. KEGG)
- **Genomas** de diversas espécies (e.g. genoma humano – Ensembl)
- **Dados expressão genética** (e.g. Microarrays em Stanford, NCBI GEO)
- **Bibliografia** (e.g. MEDLINE, PubMed)
- **Taxonomia** (e.g. Tree of Life)
- **Ontologias** (terminologia)
- **Mutações** / **doenças genéticas** (e.g. SNPs, OMIM)

(…)

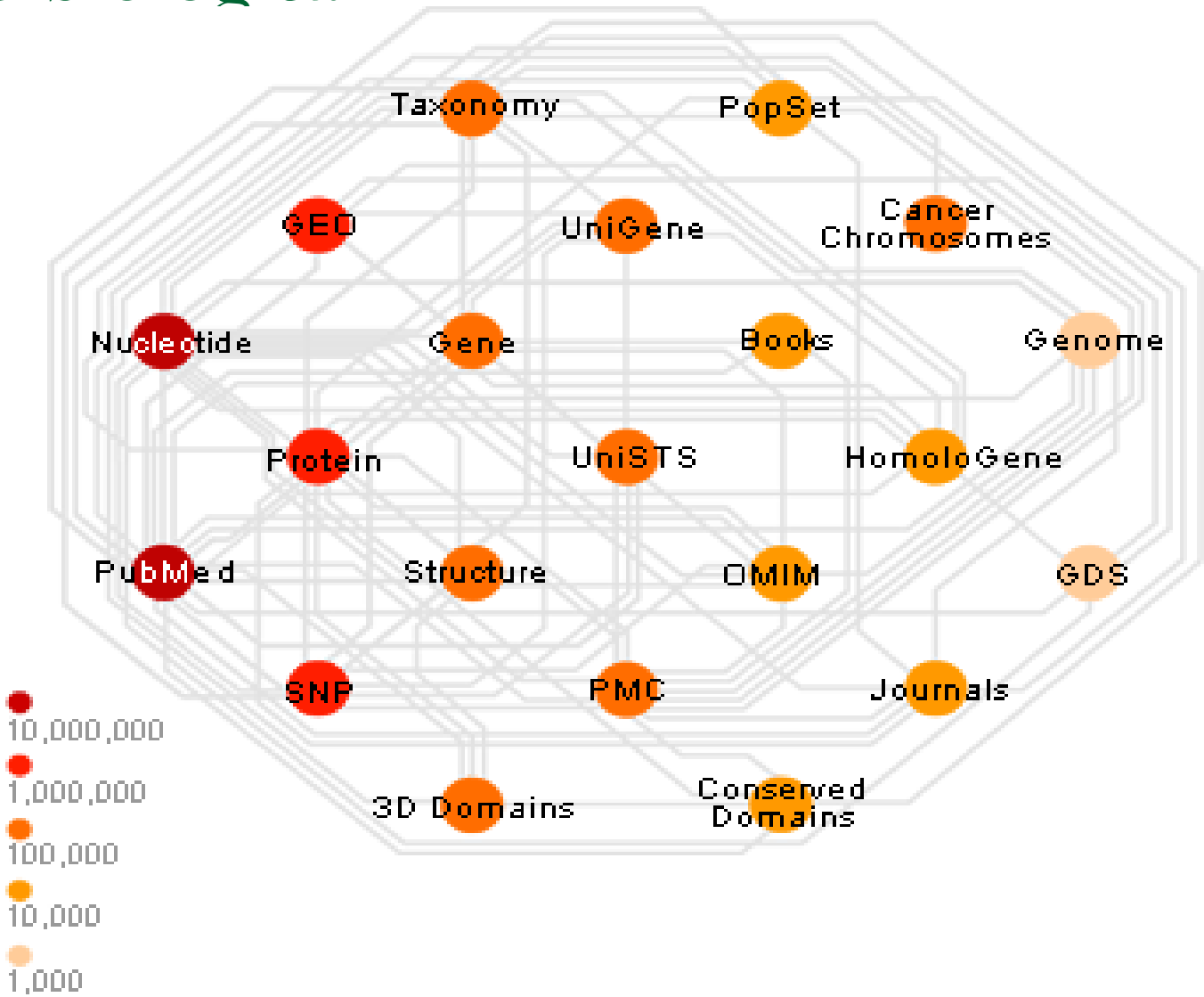
Bases de dados biológicas: tecnologias

- As primeiras gerações de bases de dados assentavam em ficheiros de texto (*flat files)*;
- Evolução na quantidade de dados e nos requisitos da análise obrigou a modelos mais complexos (e.g. modelo relacional);
- Grandes requisitos ao nível da **integração** das diversas bases de dados e aplicações.



Sistemas integrados de procura de informação biológica

- ENTREZ (NCBI)



Alinhamento/ similaridade de sequências

- Objectivo: comparar sequências de DNA ou proteínas:
  - produzindo o melhor **alinhamento**, caracter a caracter, entre duas sequências;
  - determinando a sua **similaridade**.
- Problema pode ser visto como um problema de **optimização** que dadas duas sequências e uma função de mérito, retorna o melhor alinhamento possível entre as duas sequências.

Razões para alinhar sequências

- Existem **muitas** sequências para as quais a **estrutura e a função não são conhecidas**.
- Existem **algumas** sequências para as quais a **estrutura e/ ou a função são conhecidas**.
- Um bom alinhamento de duas sequências implica que estas são similares e que **poderão ter uma ascendência comum**.
- Duas **sequências similares** têm uma probabilidade mais alta de terem **estruturas e funções semelhantes**.

Universo de procura de soluções

- Quantos possíveis alinhamentos existem entre duas sequências ?
  - Assumindo sequências ambas de tamanho **n** e que podem existir espaçamentos.
  - Número total de hipóteses:

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Para **n** = 20 – número de hipóteses cerca de 120 biliões !!!!

Funções de mérito

- Tipicamente a função de mérito adoptada é **aditiva**, i.e. corresponde à soma dos termos associados a cada par de caracteres ou espaçamentos, sendo definida a partir de:
  - **Matriz de substituição** para a co-ocorrência de caracteres;
  - Função de **penalização** para a ocorrência de espaçamentos.
- A escolha destes parâmetros influencia fortemente o resultado do alinhamento.

Exemplo de avaliação de um alinhamento

```
L G P S - G C A S G I W T K S A
  | | |   |     |   | | | | |
T G P S G G - - S R I W T K S G
```

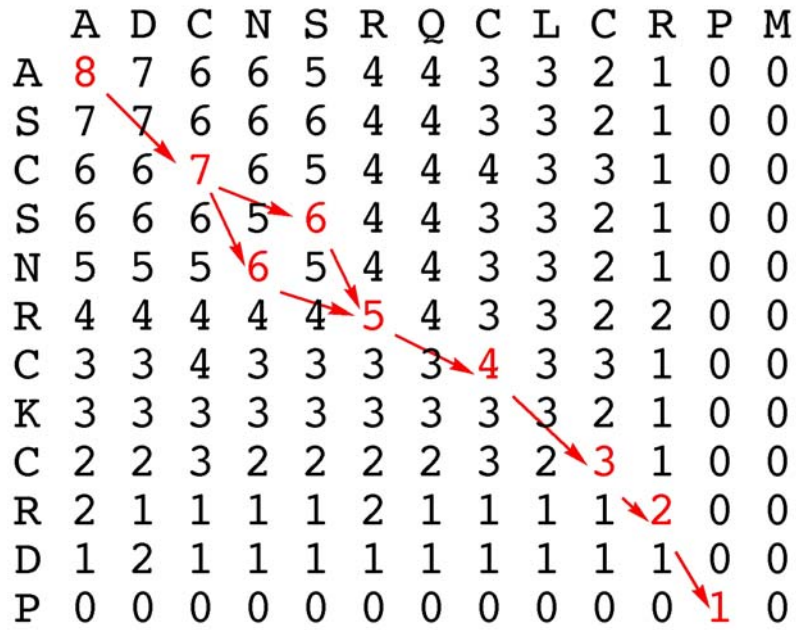
Matriz: BLOSUM62
Penalizações: g= -12; r =-2

Função de mérito do alinhamento:

-1 + 6 + 7 + 4 -12 + 6 -12 - 2 + 4 - 2 + 4 + 11 + 5 + 5 + 6 + 0 = 9

Para duas sequências: programação dinâmica

- No caso de termos duas sequências, existem algoritmos eficientes - $O(n^2)$ - para resolver o problema, baseado em **programação dinâmica** (algoritmos de *Needleman-Wunsch* e *Smith-Waterman*)



	A	D	C	N	S	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
S	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
S	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Procura em bases de dados

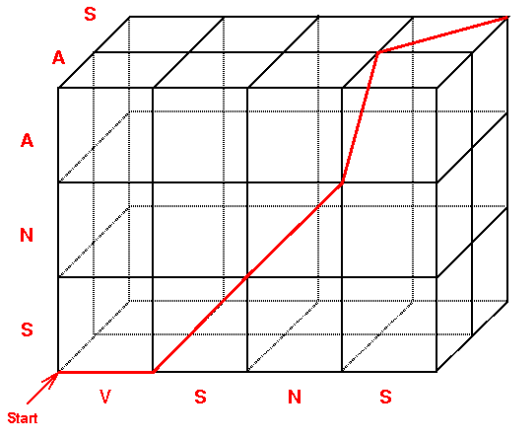
- Mas se o objectivo é **procurar sequências similares** a uma sequência alvo, em bases de dados com milhões de sequências … algoritmos de PD são lentos !!
- Solução: algoritmos heurísticos (não garantem a solução óptima) mais rápidos (cerca de 50 a 100 vezes):
  - FASTA
  - **BLAST**

Alinhamento múltiplo

- Porquê alinhar várias sequências:
  - Projectos de **sequenciação de genomas** – sequenciam-se vários segmentos cuja ordem é desconhecida e usa-se o AM para dar a ordem a estes segmentos;
  - Derivação de informação **filogenética** a partir das sequências;
  - Identificar **zonas conservadas de proteínas** – prováveis zonas activas;
  - Prever **função / estrutura** de proteínas.

Alinhamento múltiplo: um desafio

- Alinhar várias sequências – problema mais complexo: PD torna-se incomportável !!



V S N _ S
_ S N A _
_ _ _ A S

- Necessários métodos mais eficientes de optimizaçao – algumas alternativas:
  - **Progressivos** – iniciam com 2 seqs e vão adicionando as restantes;
  - **Iterativos** – consideram um alinhamento inicial que vai sendo melhorado;
  - **Estatísticos** – baseados em modelos probabilísticos.

Procura de "motifs"

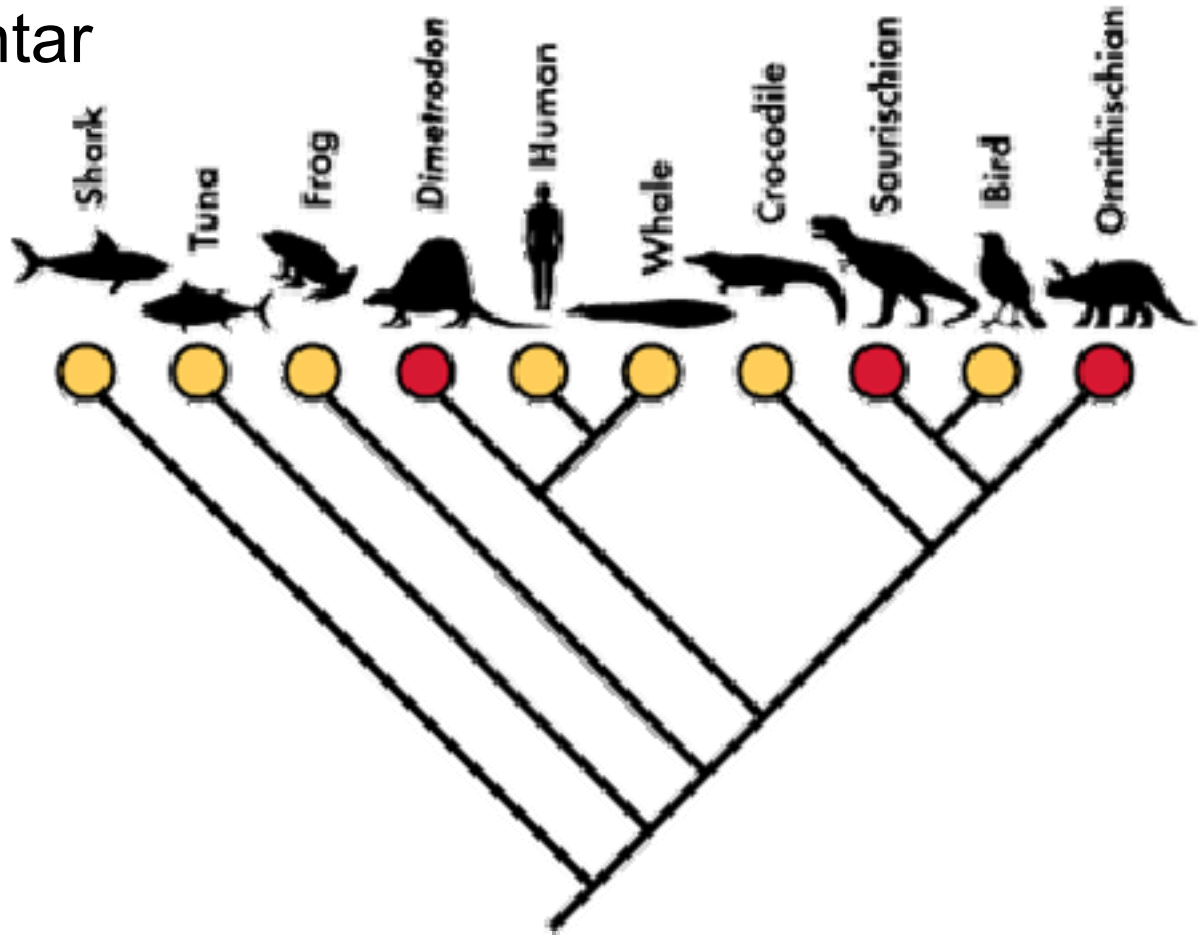
- Problema relacionado com AM: descobrir zonas (curtas) de proteínas ou DNA muito semelhantes (motifs)
  - Podem corresponder a zonas conservadas de proteínas ou a locais de regulação do DNA
- Bastante usados modelos probabilísticos, e.g. Hidden Markov Models
- Alternativa popular: algoritmos estocásticos – EM, Gibbs sampling, Algoritmos Evolucionários

Análise filogenética

- **Análise filogenética** de um conjunto de sequências (DNA, RNA, proteínas) é a determinação de como cada sequência pode ter sido derivada ao longo do processo de **evolução** natural.
- Relações evolutivas são visualizadas colocando as sequências como folhas de uma **árvore evolucionária**, onde as ramificações representam eventos de mutação (substituição, inserção, remoção).

Análise filogenética

- Pode representar relações entre espécies:



Análise filogenética: aplicações

- Determinar a **árvore da vida**, ou seja, a evolução das diferentes espécies auxiliando os métodos tradicionais baseados na morfologia;
- Ajuda na determinação da **função** de sequências de DNA/ proteínas;
- Análise de espécies com mutações rápidas (e.g. virus) – pode ajudar na **epidemiologia**;

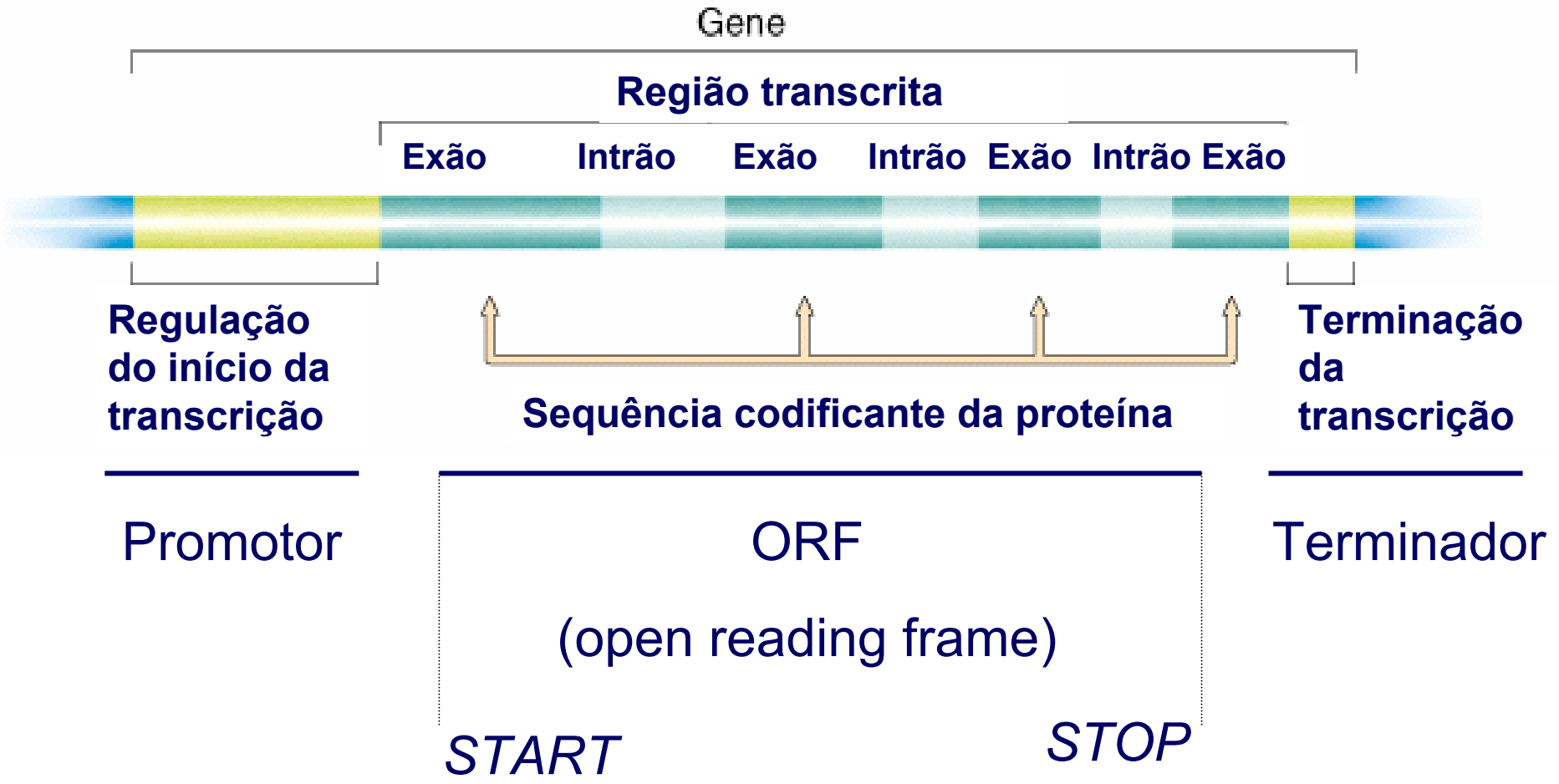
Algoritmos de análise filogenética

- Objectivo: a partir de um **conjunto de sequências** (DNA ou proteínas) determinar a **árvore evolucionária** que melhor explique a sua evolução.
- Problema de **optimização**: de entre todas as árvores possíveis, escolher a que maximiza uma dada função objectivo.
- Espaço de procura tipicamente bastante grande – problema complexo.

Algoritmos de análise filogenética

- **Máxima parcimónia** (ou mínima evolução)
  - Retornam a árvore que minimiza nº de mutações necessárias para explicar a variação das seqs.
- Baseados na **distância**
  - Baseia-se na distância (alterações) entre pares de sequências
- **Máxima verosimilhança**
  - Emprega modelos probabilísticos

Splicing: mais complexidade …



Previsão de genes (zonas codificantes)

- Desenvolvimento dos projectos de **sequenciação de genomas** fez disparar a quantidade de sequências de DNA, cuja função é desconhecida.
- Papel de algoritmos capazes de identificar zonas de codificação (de proteínas, de RNA) e de controlo da expressão genética foi reforçado.
- **Algoritmos de previsão automática de genes**, em combinação com pesquisas em BDs de sequências com funções conhecidas, são ferramentas primordiais na **anotação** dos genomas.

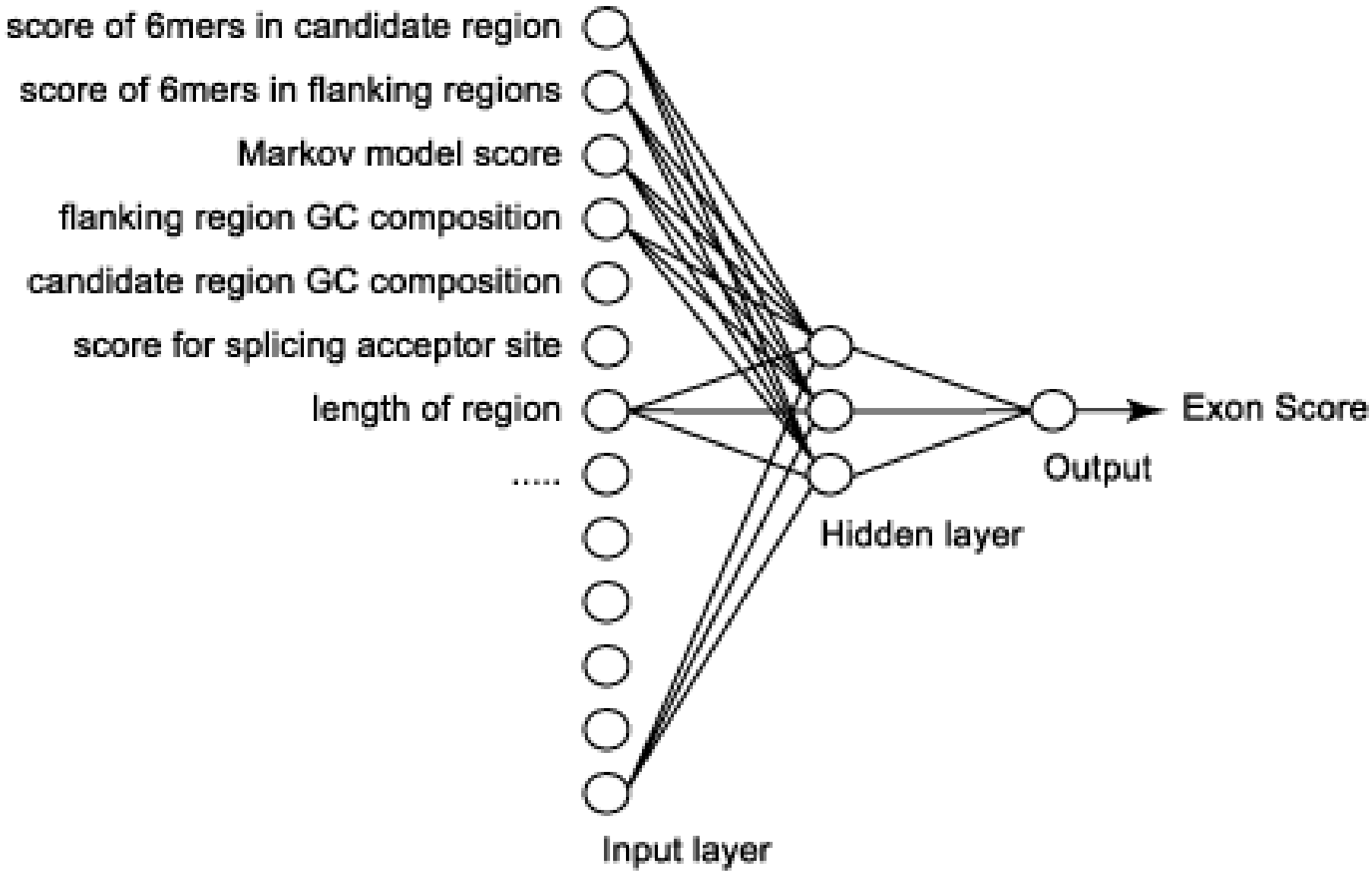
Previsão de genes

- Métodos baseados na **homologia**
  - Métodos mais simples de procurar genes:
  - Pesquisam sequências semelhantes em outras espécies, ou na mesma espécie;
- Testes estatísticos
  - **Distribuição estatística dos codões** nas zonas que codificam genes podem apresentar características diversas do restante DNA (não codificante).
  - Estas diferenças podem ser usadas para criar **testes que possam atribuir probabilidades de dadas zonas poderem conter genes**.

Em eucariotas como nós … o problema é complexo (dado o splicing …)

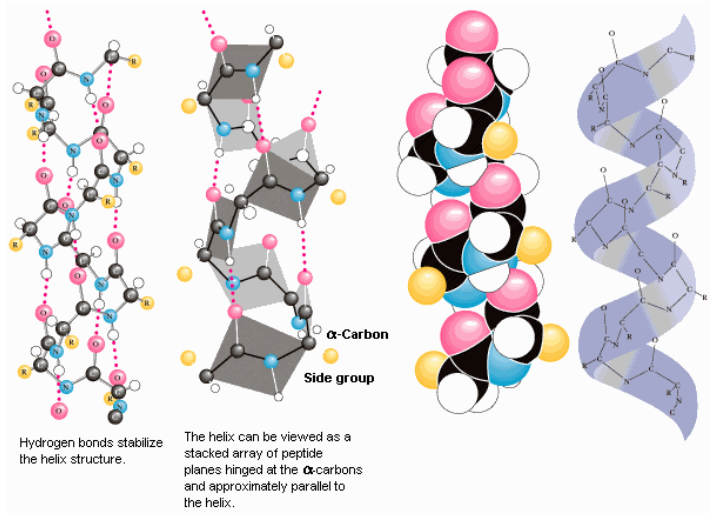
- Métodos mais usados – **Aprendizagem Supervisionada**:
  - **Treinar** um modelo de discriminação dos exons, em relação a zonas não codificantes, usando sequências cuja função de cada zona é conhecida (generalizando).
  - **Aplicar esse modelo a novas sequências** cuja função é desconhecida.
  - Tipicamente, modelos treinados numa espécie não podem ser aplicados a outras espécies.

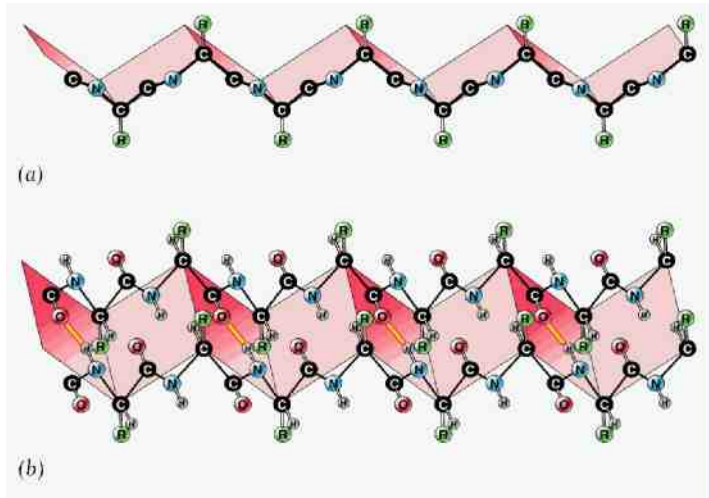
RNA do sistema GRAIL –previsão de genes



Da sequência à estrutura de proteínas

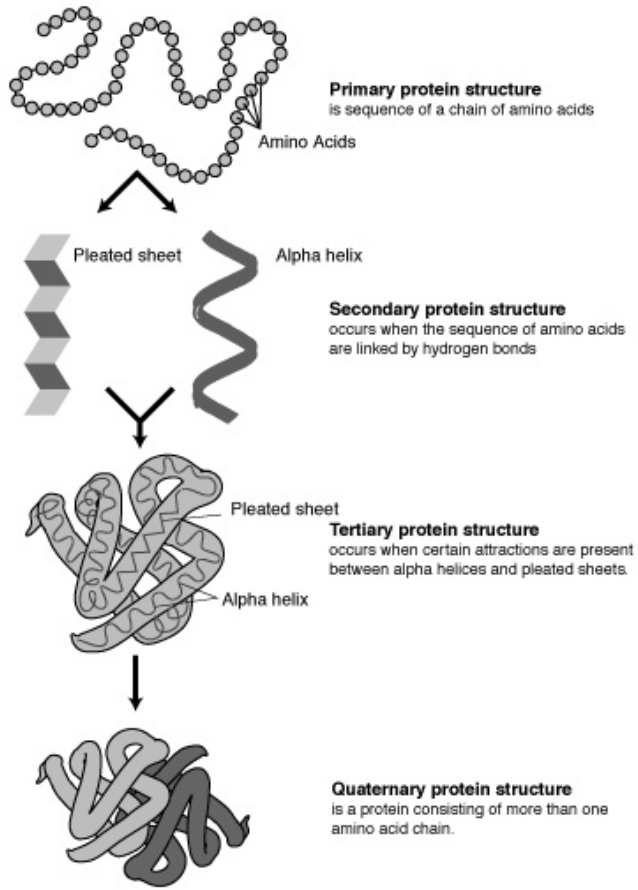
- As proteínas tendem a enrolar-se para um estado 3D de mínima energia.
- O processo de enrolamento começa enquanto a tradução está a ser realizada.
- Os resíduos hidrofóbicos são “enterrados” no interior da estrutura formando α-hélices.
- A maior parte das proteínas tomam a forma de estruturas secundárias: α-hélices e β-sheets.





A importância da estrutura das proteínas

- A estrutura que uma proteína adopta é **vital para a sua função bioquímica**
- A estrutura determina quais dos seus aminoácidos estão expostos e levam a cabo a sua função
- A sua estrutura determina com que produtos pode reagir



Estrutura e funções: proteínas

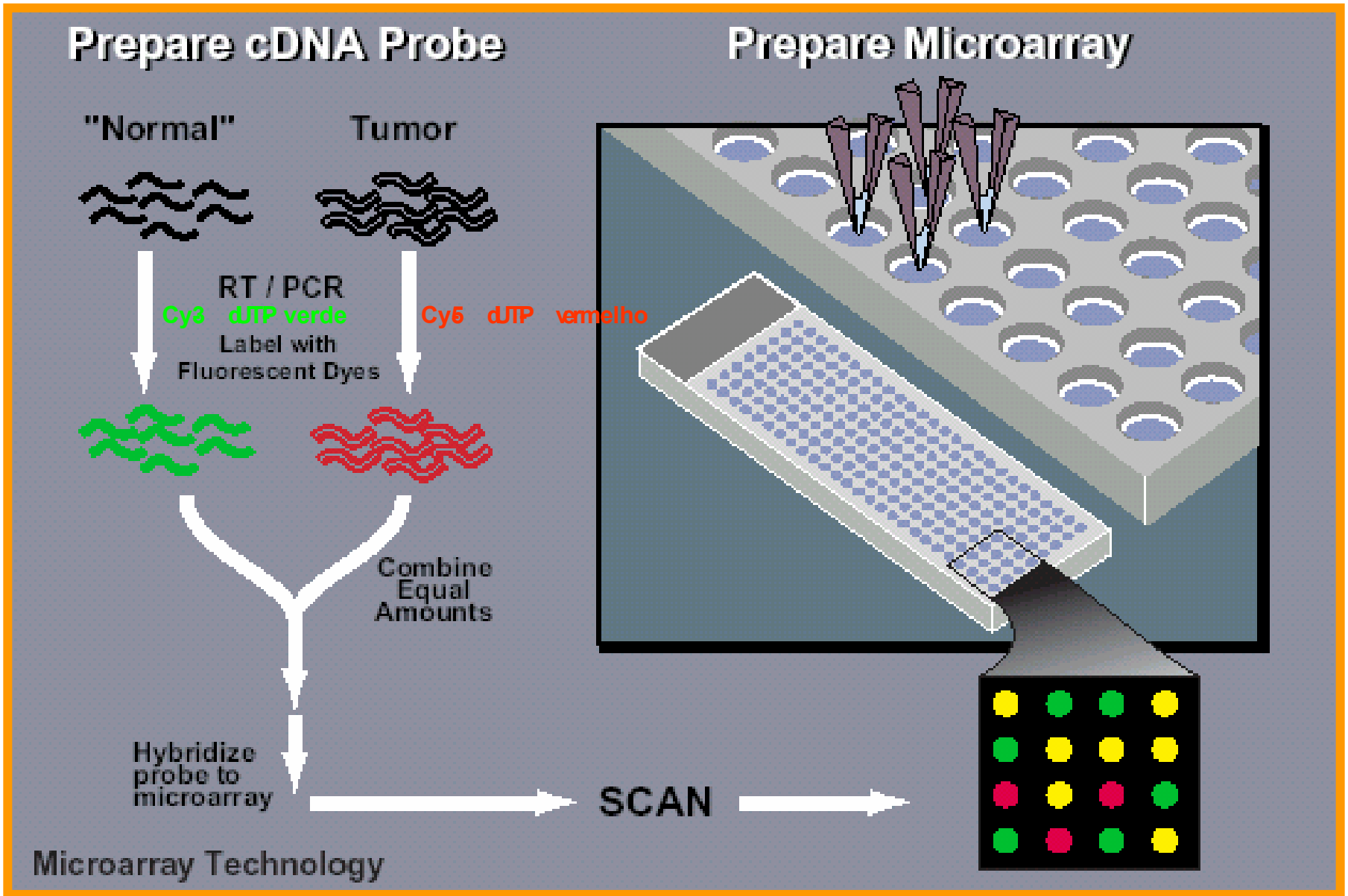
- Perceber a função de uma proteína dada apenas a sequência de AAs é um problema muito difícil.
- A própria determinação da estrutura da proteína a partir da sua sequência de AAs é um problema em aberto (o chamado ***Santo Graal da BioInformática***), que depende de muitas variáveis.
- Abordagens actuais tipicamente procuram proteínas similares e trabalham por analogia.

Inferência da função dos genes

- Um dos grandes objectivos da Bioinformática é a descoberta da função associada a cada gene (**genómica funcional**);
- Quando temos um novo gene, a comparação da sua sequência com sequências conhecidas pode ajudar na descoberta da função, mas nem sempre este método funciona (em cerca de 40% dos casos);
- **Microarrays** – nova técnica que permite aos biólogos inferir a função de um gene a partir de dados respeitantes à sua expressão em diversas condições

Microarrays de DNA

= o estudo de milhares de genes em simultâneo =



Dados de Microarrays

- Dados de microarrays são normalmente transformados numa matriz de intensidades
- A matriz de intensidade permite que os biólogos cheguem a correlações entre genes diferentes e que tentem perceber como as suas funções podem estar ligadas
- O clustering ajuda a chegar a estes resultados

Intensidade (nível de expressão) do gene na condição X

Condição	X	Y	Z
Gene 1	10	8	10
Gene 2	10	0	9
Gene 3	4	8.6	3
Gene 4	7	8	3
Gene 5	1	2	3

Clustering de dados de Microarrays

- Cada gene (linha da matriz) é encarado como um ponto num espaço N-dimensional
- Criar uma **matriz de distâncias** entre cada par de genes (necessário usar uma dada métrica para calcular a distância – e.g. euclideana)
- Pares de genes com distâncias pequenas partilham os mesmos padrões de expressão, o que pode indiciar funcionalidades similares ou relacionadas
- Clustering revela grupos de genes com padrões de expressão semelhante, logo **potencialmente relacionados funcionalmente**.

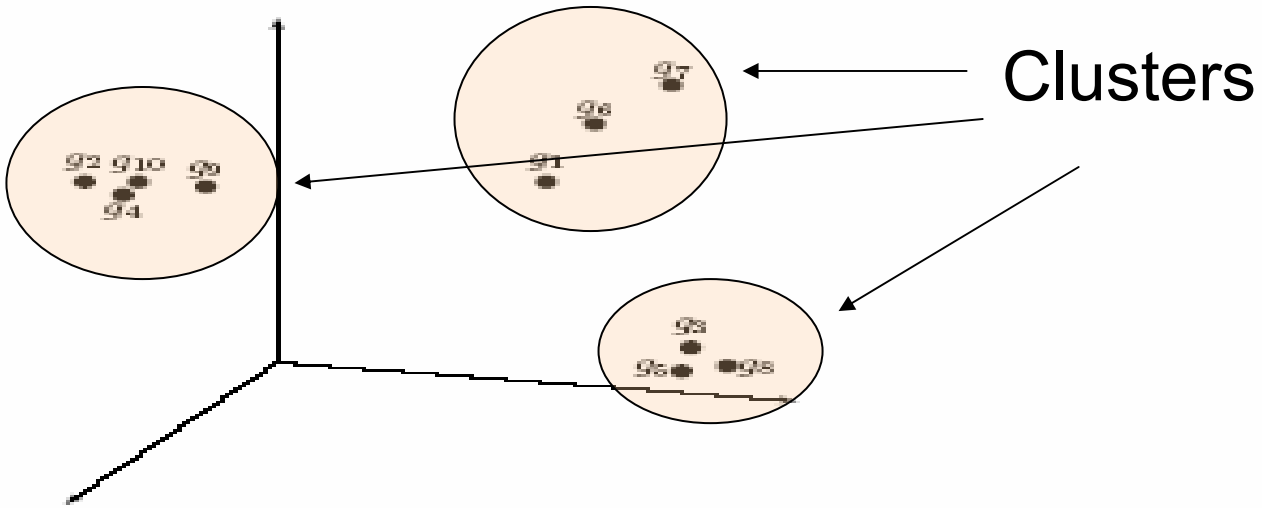
Clustering de dados de Microarrays (exemp)

Time	1 hr	2 hr	3 hr
g_1	10.0	8.0	10.0
g_2	10.0	0.0	9.0
g_3	4.0	8.5	3.0
g_4	9.5	0.5	8.5
g_5	4.5	8.5	2.5
g_6	10.5	9.0	12.0
g_7	5.0	8.5	11.0
g_8	2.7	8.7	2.0
g_9	9.7	2.0	9.0
g_{10}	10.2	1.0	9.2

(a) Intensity matrix, I

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

(b) Distance matrix, d



(c) Expression patterns as points in three-dimensional space.

Análise de dados de microarrays: o futuro

- Integração dos dados de microarrays com conhecimento adquirido sobre o genoma, a expressão dos genes e os mecanismos de regulação no organismo em estudo;
- Extracção automática de redes metabólicas e de regulação genética a partir de dados de microarrays e outros métodos experimentais.
- Uso de métodos de aprendizagem supervisionada para classificação automática dos genes e das suas funções.

Biologia de Sistemas: um dos rumos para o futuro

- Objectivo:
  - Criar modelos que permitam prever *in silico* o comportamento das células em qualquer situação.
- Como ?
  - Tirando partido dos “novos” dados experimentais de medição de vários tipos de moléculas na célula (e.g. microarrays, proteómica, etc)
  - Usando ferramentas computacionais para criar e simular modelos das reacções metabólicas e mecanismos de regulação.

Bioinformática: outros desafios do futuro

- Grandes desafios actuais ao nível pós-genómico:
  - Previsão automática da estrutura das proteínas a partir da sequência;
  - Determinação da função de cada gene: genómica funcional
  - Mecanismos de expressão e regulação: determinação automática de redes genéticas
- O grande desafio (utópico?) da Bioinformática:
  - Um modelo completo de um ser vivo (unicelular para começar !!)