Número 1-3 Volume XXII (XCVII) Lisboa 2001

brotéria genética

ÓRGÃO DA SOCIEDADE PORTUGUESA DE GENÉTICA

BROTÉRIA GENÉTICA, Lisboa, XXII (XCVII) 81-91 2001

BIOINFORMÁTICA: UMA LISTA DE RECURSOS DISPONÍVEIS

JOÃO BORGA [1], RICARDO CORREIA [2], ALTINO CHOUPINA [3]
[1] Escola Superior Agrária de Bragança, Apartado 172 5300 Bragança, Portugal,
Tel.: 00351-96-2878246, Fax: 00351-273-325405, e-mail: borga@excite.com
[2] Escola Superior Agrária de Bragança, Apartado 172 5300 Bragança, Portugal,
Tel.: 00351-96-6852346, Fax: 00351-273-325405, e-mail: rscorreia@portugalmail.com
[3] Escola Superior Agrária de Bragança, Apartado 172 5300 Bragança, Portugal,
Tel.: 00351-273-303371, Fax: 00351-273-325405, e-mail: albracho@ipb.pt

INTRODUÇÃO

Enquanto temos a compreensão básica do funcionamento do gene quando codifica sequências de proteínas especificas, sente-se a falta de informação relativa ao papel que o ADN tem em doenças especificas ou nas funções de milhares de proteínas que são produzidas. Os métodos utilizados na recolha, armazenamento, identificação, análise e correlação desta imensa e complexa informação, estão reunidos numa área científica designada por bioinformática. Todo esse trabalho produz um «oceano» de informação que só pode ser «navegado» com a ajuda de métodos computorizados. O objectivo desta área é dotar os cientistas com os meios certos para explicar:

- processos biológicos normais;
- disfunções desses processos que originem doenças;
- abordagens que permitam a descoberta de novas curas.

O conhecimento derivado das tecnologias genómicas e computacionais aumenta em progressão geométrica. A compreensão dessa avalanche de dados está intimamente vinculada ao formidável desenvolvimento na área da bioinfor-

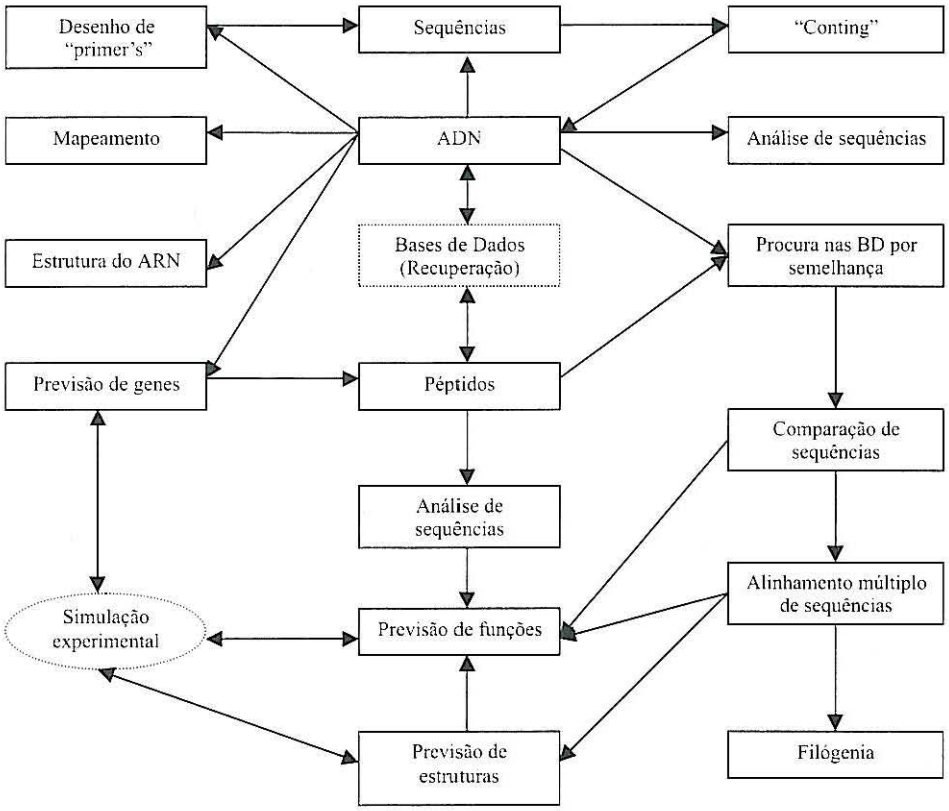
mática. Ao possibilitar a avaliação global dessa extraordinária quantidade de dados, a bioinformática tem acelerado consideravelmente as descobertas científicas.

Este crescimento tem como consequência uma grande oferta de produtos, serviços e informação, de tal forma que manter-se actualizado, localizar e utilizar as últimas novidades tornou-se uma actividade de tempo inteiro.

A bioinformática como área cientifica recolhe técnicas e ferramentas de três disciplinas:

- biologia molecular, fonte da informação a analisar;
- informática ou ciência computacional, providencia o hardware para a análise e as redes para partilhar os resultados;
- matemática, origem dos algoritmos utilizados na análise da informação.

Na inter-relação das três áreas, atrás referidas, ficam criadas as bases para a aplicação da bioinformática na biologia molecular, como se pode observar no seguinte esquema:



Esquema 1: Aplicações da bioinformática na biologia molecular (Walker, 2000)

A intenção deste artigo, e de outros disponíveis na rede, é de compilar uma lista de ferramentas e recursos informáticos utilizados pelos cientistas no tratamento de informação proveniente da sequenciação. Embora inicialmente tenhamos tentado criar um perfil completo de todos os recursos disponíveis, rapidamente foi evidente o dinamismo e constante actualização deste campo o qual suplantou esse objectivo. Propomos então as referências de base, as mais completas e as últimas novidades.

Criamos então uma divisão em quatro categorias: software de análise de sequências, software de predição de estruturas proteicas, servidores de recursos «on-line» e por último deixamos uma lista de locais de interesse na Internet que podem abreviar o tempo de pesquisa. Optamos pela selecção destas categorias por considerarmos que analisam de uma forma compreensiva o dogma da biologia molecular.

ANÁLISE DE SEQUÊNCIAS DE NUCLEÓTIDOS

No software de análise do genoma podem encontrar-se diversos pacotes de programas, os quais acompanham todo o processo desde a recepção dos gráficos provenientes do sequenciador até à publicação dos dados em bases de dados «on-line». Estas características, juntamente com o acesso grátis para académicos, a compatibilidade de ficheiros, e a sua data de concepção são os principais factores de selecção nas escolhas realizadas.

Destacamos que muitos dos serviços realizados por estes programas são também realizados por alguns programas disponibilizados «on-line», tendo estes a desvantagem de em cada consulta necessitarem de uma ligação à rede, não estando estes no PC ou MAC que hoje em dia se tornou indispensável no laboratório, mas tendo como vantagem o facto de estes recursos on-line serem actualizados regularmente.

Staden package (http://genome.wustl.edu/gsc/new/staden/staden_home.html) – Pacote de programas bastante completo no âmbito da análise de sequências de nucleótidos, é gratuito para estudantes e para investigadores, permitindo a requisição via correio ou directamente da rede. Este pacote contém os seguintes programas:

- **Gap4** – Este programa é a ferramenta principal deste pacote, realiza a compilação, junção de sequências, rectificação da compilação, lê pares de sequências e permite a edição das mesmas (Fig. 1);

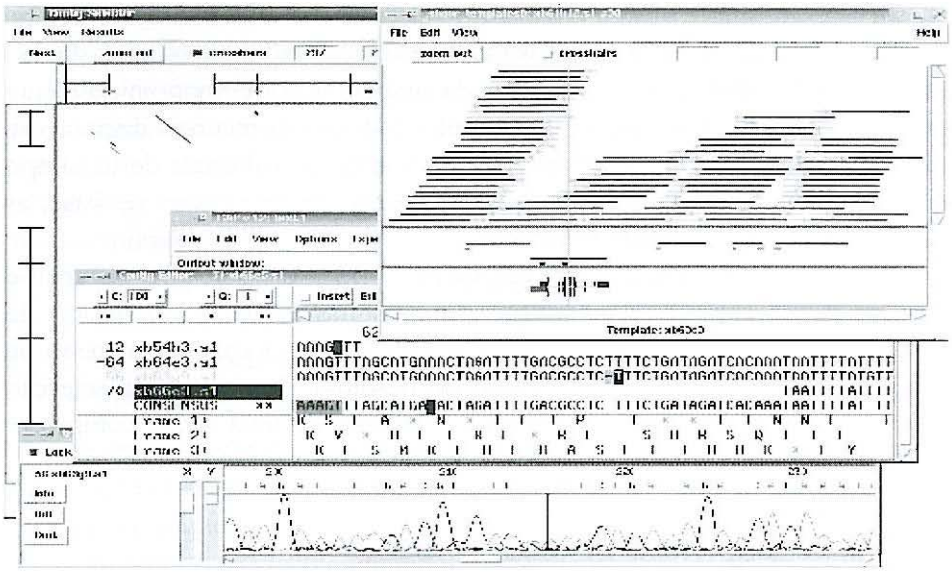


Figura 1: Interface do Gap4 (Manual do Staden Package)

- **Pregap4** – Permite uma análise da informação, recepção de informação directamente do sequenciador em vários formatos. É de certa forma, a porta de entrada para este pacote de programas;
- **Trev** – Rápido e eficaz, permite o visionamento de sequências em formatos ABI, ALF ou SRF;
- **Trace diff** – Realiza automaticamente a localização de pontos de mutação comparando a sequência com as sequências de referência. Suporta qualquer número de sequências e permite a visualização de resultados pelo gap4;
- **Sip4** – Compara pares de sequências de diversas formas, apresentando muitas vezes os resultados graficamente. Permite a comparação, base com base, proteína com proteína e proteína com base;
- **Nip4** – Analisa sequências de nucleótidos para encontrar genes, locais de restrição; permite a tradução, etc.

Dnatools (www.dnatools.dk) – Outra proposta de pacote para PC é o *Dnatools*. Constitui um concorrente à altura da anterior referência, com uma actualização feita recentemente. Destaque-se, ainda, uma actualização da biblioteca de enzimas de restrição «rebase» que data do início de 2001 e está em constante actualização. Neste pacote estão contidos os seguintes programas:

- **Clustral** – Aplicação que permite alinhamento de várias sequências e a sua manipulação. Este programa foi uma edição que por si só, permite a sua utilização sem o resto do pacote;
- **Blastall, Formatdb** – Permitem o acesso a bases de dados nos cinco programas blast;
- **Blastel3** – Juntamente com o pacote vem também esta opção, que realiza o mesmo trabalho que o anterior mas com a vantagem de ser completamente compatível com procuras no ***NCBI***;
- **Convert trace** – Permite ao *dnatools* importar e exportar ficheiros convertendo os cromatogramas provenientes dos mais comuns sequenciadores;
- **Chromas** – Este programa permite também a visualização de ficheiros provenientes de sequenciadores.

pDRAW 32 – Um programa para ser utilizado numa plataforma Windows, com uma interface agradável e intuitiva, estando disponível gratuitamente na Internet no site: (http://www.crosswinds.net/~acaclone/).

Com este programa é possível realizar várias operações, tais como: anotações relativas ao ADN em estudo, clonar ADN, editar sequências, analisar sequências, seleccionar enzimas, exportar gráficos e texto, calcular a temperatura óptima para PCR, calcular homologias entre dois fragmentos de ADN e ainda um ficheiro de ajuda científico.

DNASTAR™ – Outro pacote informático que tem vindo a ter grande utilização é o DNASTAR™ (cópia de demonstração enviada por Dnastar, Ltd. Abacus House, Manor Rd. West Ealing London W130AS, Reino Unido) a qual possui programas com os quais se pode fazer a edição básica de sequências, comparação de sequências e algumas características físico-químicas bem como a construção de plasmídeos, etc.

PREVISÃO E VISUALIZAÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS

A previsão de estruturas proteicas é um campo de pesquisa actual, sendo o software apresentado uma ferramenta complementar, mais do que uma ferramenta exacta ou definitiva. Isto porque o software baseia-se em predicção por comparação com outras proteínas já analisadas e com modelos construí-

dos a partir destes estudos. No entanto para um trabalho definitivo existem na rede, imensos servidores e laboratórios que estão referidos nos recursos «on-line» e que apresentam resultados exactos e alcançados de forma experimental. Deixamos também a advertência, que este tipo de software é ávido em memória do computador já que alguns geram imagens a três dimensões e a cores.

Fica também implícita a possibilidade de depois de estudar um gene e com pouco esforço, seguir o caminho da sua tradução em aminoácidos e conseguir de uma forma aproximada prever a estrutura proteica a que este dá origem.

Swiss-PdbViewer (http://ca.expasy.org/spdbv/) – Um programa disponibilizado para MAC, PC, SGI e LINUX de forma gratuita. É uma aplicação de relação amigável com o utilizador, que permite analisar ao mesmo tempo diversas proteínas. As proteínas podem ser sobrepostas a fim de deduzir alinhamentos estruturais e comparar os seus locais activos e/ou todas as outras informações relevantes; mutações de aminoácidos, pontes de hidrogénio, ângulos e distâncias entre átomos são fáceis de obter graças ao menu e gráfico intuitivo. Além disso, o *Swiss-PdbViewer* está interligado à **Swiss-Model**, um servidor automatizado de modelos de homologia, desenvolvido na *Glaxo Welcome Experimental Research*, em Genebra. Trabalhar com estas duas aplicações reduz, de forma considerável, a quantidade de trabalho necessária para gerar modelos, pois é possível colocar uma sequência primária da proteína num molde 3D e obter uma resposta imediata de como a proteína será aceite pela estrutura referida, antes de submeter um pedido para construir ciclos em falta e refinar as cadeias laterais.

GOpenMol (http://www.csc.fi/~laaksone/gopenmol/distribute/) – O GOpenMol é um interface gráfico para o estudo de proteínas, este software também pode ser utilizado de acordo com os autores para a análise e visionamento de trajectórias de dinâmicas moleculares, visionamento de orbitais moleculares, densidade electrónica, potenciais electrostáticos, a partir de programas como ***Pc gamess*** e ***Jaguar***. A versão actual data de Março de 2001. Utiliza como sistemas operativos o Linux, Windows e SGI IRIX. Constitui uma novidade com capacidade de aceitar os mais diversos formatos de ficheiros constituindo por si só a grande capacidade deste software. Os gráficos apresentados, na figura 2, são produzidos por este programa e representam bem as possibilidades do mesmo.

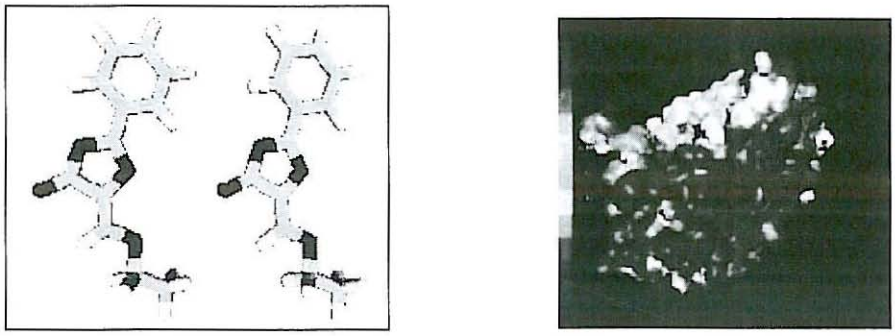


Figura 2: Análise de uma estrutura tridimensional e cálculo do potencial isoeléctrico, com o GOpenMol (Adaptado de http://www.csc.fi/~laaksone/gopenmol/distribute/)

Cn3d (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) – Este é um programa de ajuda à visualização tridimensional das estruturas proteicas, especificamente desenhado para permitir a visualização de ficheiros provenientes do ***NCBI-Entrez***, este software, utiliza como plataformas o Windows, MAC-OS e Unix. Este programa permite a visualização em simultâneo da estrutura, sequência e alinhamento. Com uma actualização a meio do ano passado, esta permitiu o melhoramento dos gráficos, assim como do visualizador do alinhamento de sequências, coloração de sequências de conservação e possibilidade de gravar as especificações de análise num ficheiro.

SERVIDORES DE RECURSOS ON-LINE

Aplicações

Na Internet têm-se tornado disponíveis, nos últimos anos uma grande variedade de programas «on-line», estes programas não necessitam de instalação no computador cliente e têm capacidades muitas vezes superiores às que conseguimos alcançar com os micro-computadores que utilizamos habitualmente. Na maioria dos casos, as capacidades destes servidores são postas à disposição da comunidade, sendo as verdadeiras mais valias tiradas em aplicações que requerem grandes capacidades de processamento, como a identificação de homologias ou construção gráfica de estruturas 3-D. Na lista abaixo indicada ficam algumas das aplicações que considerámos mais interessantes, alertamos no entanto que os recursos referidos são, apenas, uma amostragem do vastíssimo leque de aplicações disponíveis.

Homologia de sequências (Sequence Homology):

BLAST	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast?Jform=0
PSI-BLAST	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast
FASTA	http://www2.ebi.ac.uk/fasta3/
HMM	http://pfam.wustl.edu/hmmsearch.shtml
Smith-Waterman	http://crick.genes.nig.ac.jp
ClustalW	http://pbil.ibcp.fr/NPSA/npsa_clustalw.html
RepeatMasker	http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
Translate	http://arbl.cvmbs.colostate.edu/molkit/translate
Signal	http://www.cbs.dtu.dk/services/SignalP/index.html=submission
Transmembrane	http://www.ch.embnet.org/software/TMPRED_form.html
BLOCKS search	http://blocks.fhcrc.org/blocks/blocks_search.html
BLAST-Prodom	http://www.toulouse.inra.fr/prodom/doc/blast_form.html

Previsão de genes (Gene-Prediction):

GENSCAN	http://gnomic.stanford.edu/
FGENEH	http://searchlauncher.bcm.tmc.edu/
GeneID	http://www1.imim.es/geneid.html
Genie	http://www.fruitfly.org/seq_tools/genie.html
GRAIL	http://compbio.ornl.gov/Grail-1.3/
SpliceView	http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html
GeneBuilder	http://l25.itba.mi.cnr.it/~webgene/genebuilder.html
CpG Island	http://l25.itba.mi.cnr.it/genebin/wwwcpg.pl
TATA Signal	http://l25.itba.mi.cnr.it/~webgene/wwwHC_tata.html
GenView	http://l25.itba.mi.cnr.it/~webgene/wwwgene.html
HCPolyA	http://l25.itba.mi.cnr.it/~webgene/wwwHC_polya.html
GeneFinder	http://sciclio.cshl.org/genefinder/human.htm
AAT	http://genome.cs.mtu.edu/aat.html
GRAIL	http://compbio.ornl.gov/Grail-1.3/
SpliceView	http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html
GeneBuilder	http://l25.itba.mi.cnr.it/~webgene/genebuilder.html
CpG Island	http://l25.itba.mi.cnr.it/genebin/wwwcpg.pl
TATA Signal	http://l25.itba.mi.cnr.it/~webgene/wwwHC_tata.html
GenView	http://l25.itba.mi.cnr.it/~webgene/wwwgene.html
HCPolyA	http://l25.itba.mi.cnr.it/~webgene/wwwHC_polya.html
GeneFinder	http://sciclio.cshl.org/genefinder/human.htm
AAT	http://genome.cs.mtu.edu/aat.html

Previsão de estruturas (Structure-Prediction):

SSP-BCM	http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html
GOR-SSP	http://absalpha.dcrt.nih.gov:8008/gor.html
RNA-SSP	http://www.genebee.msu.su/services/rna2_reduced.html
ICM-SSP	http://24.3.130.175:7788/serv/wpredictss.htm
DSC-SSP	http://bonsai.lif.icnet.uk/bmm/dsc/dsc_form_align.html
Predator	http://www.embl-heidelberg.de/cgi/predator_serv.pl
Coiled-coil	http://www.isrec.isb-sib.ch/software/COILS_form.html
FromAlignment	http://kestrel.ludwig.ucl.ac.uk/zpred.html

Previsão de dobras (Fold-Prediction):

UCLA-threading	http://www.doe-mbi.ucla.edu/people/fischer/TEST/getsequence.html
123D-threading	http://www-lmmb.ncifcrf.gov/~nicka/run123D.html
H3P2	http://ampere.doe-mbi.ucla.edu:8805/submit2.html

BASES DE DADOS

O armazenamento da informação é um factor importante para a bio-informática, são necessários padrões que permitam um armazenamento da informação e das suas anotações de forma a permitir um rápido acesso e a partilha desta entre as várias bases de dados sem problemas de compatibilidade, quer estas sejam de sequências, genomas, proteínas ou enzimas. As grandes bases de dados permitem uma partilha de dados, entre si, aumentando assim a rentabilização dos recursos e consequentemente permitindo que toda a comunidade a consulte e se mantenha actualizada.

No caso das bases de dados de nucleótidos estas são formatadas e inscritas numa das quatro grandes bases de dados as quais partilham entre si as sequências, estas estão organizadas de acordo com a sua proveniência taxonómica e tipo de sequência, identificada por uma sigla de três letras, EST (*expressed sequence tags*), STS (*sequence tagged sites*), GSS (*genomic survey sequences*).

Estas bases de dados são da responsabilidade de instituições governamentais e sem fins lucrativos, sendo portanto as referencias consensuais na área:

- **Genbank** - http://www.ncbi.nlm.nih.gov/
- **EMBL** - http://www.ebi.ac.uk/
- **DDBJ** - http://www.ddbj.nig.ac.jp/
- **GSDB** - http://www.ncgr.org/

As bases de dados de proteínas poderiam ser consideradas redundantes devido a serem originadas em grande parte, da tradução de sequências vindas de bases de dados. No entanto, estas têm a maior parte das vezes, melhores anotações do que as bases de dados de nucleótidos, e algumas chegam a ter anotações sobre processos experimentais relacionados com estas. A aplicação destas bases de dados insere-se na área da família dos genes e na função das proteínas. Por último, antes de apresentar uma lista de endereços, salientamos a qualidade de uma das bases de dados analisadas. A **Swissprot**, foi no nosso

entender a que mais se destacou, quer devido a qualidade das suas anotações como à quantidade existente.

SwissProt	http://www.ebi.ac.uk/	Base de dados de sequências proteicas
Pir	http://nbrfa.georgetown.Edu/	A maior base de dados de proteínas anotada.
Genpept	http://www.ncbi.nlm.nih.gov/	Tradução das sequências do GenBank
SPTR	http://www.ebi.ac.uk/	SwissProt + SPTREMBL + TREMBLNEW
OWL	http://www.bioinf.man.ac.uk/dbbrowser/OWL/	SwissProt+PIR+Genpept + NRL_3d
NCBInr	http://ncbi.nlm.nih.gov/	SwissProt + PIR + Gen pept + PDB + PRF

OUTRAS HIPERLIGAÇÕES / CONCLUSÃO

No trabalho realizado, a primeira dificuldade que encontrámos foi o excesso de informação obrigando a um dispêndio de tempo na triagem da mesma de forma a encontrar o que realmente é importante. Também encontrámos imensos recursos repetidos ou pura e simplesmente desactualizados e sem manutenção. Outra das questões com que nos deparámos foi a falta de informação em português, mesmo com o nosso grande irmão Brasil em pleno desenvolvimento na área da biotecnologia. Sentimos então a necessidade de produzir um espaço na rede desenvolvido na língua de Camões (http://bioinformatica.pt.st).

O que não encontrámos, mas que também sentimos necessidade foi a existência de manuais sobre a área, traduzidos para português e se possível disponíveis na rede. Deixamos então uma lista de hiperligações que nos foram úteis nesta área e durante a elaboração deste artigo:

- (http://www.igc.gulbenkian.pt) Site do Departamento de Bioinformática da Fundação Calouste Gulbenkian, com acesso às principais bases de dados, assim como informações sobre cursos de bioinformática;
- (http://www.biolinks.com) Site com constante actualização e muitos endereços não só nesta área;
- (http://www.biozentrum.uni-wuerzburg.de/biolinks/biolinks.html) Índice de hiperligações para biologia molecular, genoma, bioinformática, microbiologia, etc.;
- (http://www.public.iastate.edu/~pedro) Página de um prof. adjunto do Instituto Superior Técnico, uma das grandes referências no índex de hiperligações;

- (http://www.pasteur.fr/recherche/BNB/bnb-en.html) Página de notas e hiperligações da responsabilidade do famoso Instituto Pasteur;

- (http://www.lv.psu.edu/jxm57/biolinks.html) Uma lista de endereços de biologia clássica;

- (http://www.cellbiol.com/) Página de referência para quem aborda a área com um Macintosh, contem software, protocolos, hiperligações, etc.;

- (http://ibscore.dbs.umt.edu/biolinks.htm) Página com listas de páginas de hiperligações de todas ou quase todas as áreas da biologia;

- (http://www.expasy.ch/alinks.html) Página mantida pela **Sib switzerland**;

- (http://sgi.bls.umkc.edu/biolinks/reference.html) Uma página com actualização em Janeiro de 2001, bastante completa e com muitas hiperligações;

- (http://www.bioinformatica.com) Página com hiperligações para todos os grandes Institutos e aplicações de interesse.

BIBLIOGRAFIA

PEARSON, W. e L., D. (1988). Improved tools for biological sequence analysis. Proc. Nat. Acad. Sci. USA 85: 2444-2448.

WALKER, J. M. e RAPLEY, R. (2000). Molecular Biology and Biotecnology. Royal Society of Chemestry. University of Hertfordshire, Hatfield, UK. 4: 405-431.

Todos os sites referidos anteriormente!