Bioinformática Elementar

Diana Magalhães de Oliveira

Um Almanaque de Bioinformática para Iniciantes

Diana Magalhães de Oliveira (Ed.) – 1ª Edição

Maio, 2009

Renorbio Publicações – Editora UECE

55(85) 3101-9850 / www.uece.br

Unidade I

Bioinformática Dentro da Biologia Computacional

Capitulo 1 – Antecedentes

Diana Magalhães de Oliveira

•Apresentação
•Definições, Conceitos, Classificação
•Origem e Histórico da Bioinformática
•Âmbito das Ciências Genômicas e Pós-Genômicas
•Aplicações e Abordagens

1.1. APRESENTAÇÃO

A crescente demanda pela utilização da informática, ou melhor dizendo da computação, para a análise de dados gerados na pesquisa biológica tem marcado o século XXI de forma irreversível. Especialmente em alguns setores de pesquisa em bioquímica, genética e biologia moleculares, o uso da computação já é imprescindível para que se possa chegar a resultados definitivos a partir dos dados experimentais (como p.ex., aqueles de eletroforese bidimensional, espectrometria de massa, etc). A análise de genomas e proteomas, bem como os estudos relativos à relação estrutura e função de proteínas e de outras macromoléculas de interesse biológico são os setores em que isso mais se evidencia atualmente.

Não se deve esquecer que as raízes da Biologia Computacional remotam aos primórdios da genética e da biologia molecular. Fisher e Haldane, por exemplo, trouxeram a análise matemática à genética Mendeliana e o trabalho seminal de Watson e Crick não passa, essencialmente, de um típico artigo teórico de análise de dados

publicados sobre a difração de raios-X na molécula de DNA. A migração de cientistas das áreas de física e matemática para a biologia molecular, que se iniciou nos anos de 1960s e continua até hoje, tem estimulado uma gama de abordagens analíticas e computacionais aplicadas aos problemas em biologia molecular. Grandes ícones desta migração, cientistas conversores iniciais, incluem Walter Gilbert, David Botstein, Michael Waterman, e Temple Smith, dentre vários outros que ajudaram a formar a Biologia Computacional como a entendemos nos dias atuais.

Com a explosão de informações relativas a seqüências e estruturas protéicas disponíveis aos pesquisadores e ao público em geral, o campo da Bioinformática ou Biologia Computacional está cada vez mais envolvido na elucidação dos aspectos desconhecidos da estrutura e função de genes e proteínas. Evidentemente, a Bioinformática não se limita apenas a isso; em diversas áreas da biologia e da medicina podem ser usadas diferentes técnicas computacionais para resolver problemas médicos e/ou biológicos diversos (por exemplo, análise de imagens micro-, endo- e nanoscópicas, tomo- e ultrasonográficas, reconhecimento de padrões em sinais biológicos, etc). Concentraremo-nos aqui na Bioinformática clássica (partindo da análise de seqüências de genes e proteínas para se chegar à estrutura e à função) pela atual importância das pesquisas básicas e aplicadas em biologia molecular e biotecnologia. No entanto, devemos estar alertas para o fato de que a bioinformática vem alcançando novas e atraentes áreas de aplicação dentro da pesquisa biomédica, farmacêutica e agropecuária, as quais ampliam, sobremodo, o mercado de atuação do profissional especializado nesta área.

Abordaremos aqui os principais recursos e ferramentas – softwares, bancos de dados e sites de interesse na web - relativos à bioinformática. Ao longo dos dezessete capítulos que compõem o livro, veremos como se dá a aplicação dessas ferramentas a partir das seqüências de DNA, RNA e de proteínas obtidas por meio de técnicas bioquímicas e de biologia molecular. A montagem de seqüências e a comparação com seqüências já conhecidas, permitindo estabelecer relações filogenéticas e a predição de estruturas tridimensionais é um primeiro passo na compreensão de aspectos

estruturais e funcionais de moléculas de DNA, RNA e proteínas nos mais diversos seres vivos.

1.2. DEFINIÇÕES, CONCEITOS, CLASSIFICAÇÃO

Poder-se-ia dizer que a Bioinformática depende de quatro elementos indispensáveis para sua existência: 1) dos microcomputadores pessoais (PCs ou *desktops*); 2) dos sistemas operacionais (SOs) e softwares especializados; 3) da internet; e 4) da interpretação/inferência biológica. De fato, a disseminação de PCs conectados à internet permitiu o uso, por um número cada vez maior de laboratórios e pesquisadores, de recursos antes disponíveis apenas a certos centros de pesquisas equipados com recursos de supercomputação. Quer utilizando bancos de dados (BDs) situados em servidores/computadores remotos ou utilizando programas em computadores locais (geralmente obtidos -de forma gratuita ou não - através da rede), o avanço da Bioinformática depende intrinsecamente dos quatro itens acima citados.

O conceito de Bioinformática pode ser resumido como a utilização de técnicas advindas da matemática, estatística e computação para a análise de problemas de biologia. O termo ***bioinformática*** é um conceito relativamente recente, tendo aparecido corriqueiramente na literatura na década de 90. Contudo, como pode ser visto pelo breve histórico apresentado a seguir, a pesquisa em Bioinformática não é um assunto novo, sendo que os marcos iniciais da pesquisa datam da década de 1960.

1.2.1 Histórico / Origem da Bioinformática

Alguns dos principais eventos relacionados à Bioinformática ao longo do tempo. A maioria dos eventos listados ocorreu antes mesmo do termo "bioinformática" haver sido cunhado.

• 1869 – Friedrich Meischer – Isolamento de DNA a partir de núcleos de leucócitos.

• 1911 – Alfred Sturtevant – Primeiro mapa genético de localizações de vários genes da mosca das frutas.

• 1919 – Karl Ereky – Introdução do termo "biotecnologia" por um engenheiro Húngaro, segundo o qual biotecnologia envolve todos os trabalhos realizados com o auxilio de organismos vivos. Mais formalmente, a biotecnologia pode ser definida como "a aplicação de princípios científicos e de engenharia ao processamento de materiais por agentes biológicos a fim de prover produtos e serviços úteis."

• 1920 – H. Winkler – Introdução do termo **genoma** para denotar o conjunto completo de genes cromossômicos e extracromossômicos presentes num organismo, incluindo os vírus.

• 1920's – Ronald A. Fisher, JBS Haldane e Sewall Wright — demonstraram como a seleção natural opera na genética Mendeliana ao executar experimentos aliados a modelos matemáticos (precisos e sofisticados) sobre a evolução. Fisher, Haldane e Wright fundaram a genética de populações, área científica cujo principal objetivo é entender de forma quantitativa as razões evolutivas para um dos fatos mais óbvios da natureza: todas as espécies possuem variabilidade genética e essa variabilidade apresenta padrões. Apesar do seu talento para a matemática, o interesse de Fisher era primordialmente a biologia, como ele mesmo disse: "uma técnica matemática com interesse biológico é um terreno muito mais firme que uma técnica biológica com interesse matemático".

• 1926 – Thomas Hunt Morgan – Publicação da "teoria do gene" baseada na genética Mendeliana.

• 1933 – Arne Tiselius – Uma nova técnica, a eletroforese, é usada para separação de proteínas em solução.

• 1951 – Pauling e Corey – Proposta para a estrutura da alfa hélice e da folha beta-pregueada (Proc. Natl. Acad. Sci. USA, 27: 205-211,1951; Proc. Natl. Acad. Sci. USA, 37: 729-740, 1951).

• 1952 – Rosalind Franklin e Maurice Wilkins – Geração de dados experimentais de cristalografia de raios-X de DNA, fornecendo informação crucial que levou à elucidação da estrutura do DNA.

• 1953 – James Watson e Francis Crick – Proposta do modelo da dupla hélice para o DNA baseado nos dados obtidos por Franklin e Wilkins (Nature, 171: 737-738, 1953).

• 1954 – Perutz e colaboradores – Métodos de átomo pesado para resolver o problema

de fase em cristalografia de proteínas.

• 1955 – F. Sanger – Anúncio da seqüência da primeira proteína analisada, a insulina bovina.

• 1956 – Francis Crick e George Gamov – Anúncio do seqüência "Central Dogma" para explicar a síntese protéica a partir do DNA: a seqüência no DNA codifica seqüências de aminoácidos e a informação genética flui numa só direção - do DNA para o mRNA e para a proteína.

• 1958 – Jack Kilby – Construção do primeiro circuito integrado na Texas Instruments. Anúncio de formação da ARPA, a Advanced Research Projects Agency nos EUA.

• 1961 – Carl Gordon Heden – O microbiologista Sueco redefine a biotecnologia como a "produção industrial de bens e serviços processados pelo uso de organismos biológicos galgados na expertise em microbiologia, bioquímica e engenharia quimica." No entanto, a natureza da biotecnologia foi definitivamente alterada pelo desenvolvimento da tecnologia do DNA recombinante, técnicas com as quais a otimização de qualquer processo biotecnológico pode ser alcançada muito mais diretamente.

• 1962 – G.N. Ramachandran – Introdução do *Ramachandran Plot* como meio de verificar as várias conformações de polipeptídeos então conhecidos e também para desenvolver um bom parâmetro ou medida (*'yardstick'*) utilizável no exame e acesso de qualquer estrutura em geral, mas particularmente de peptídeos.

• 1965 – Margaret Dayhoff – Publicação do "*Atlas of Protein Sequence and Structure*", a primeira coleção completa de seqüências de proteínas e seus aminoácidos compiladas e que serviram de fonte para o estabelecimento das matrizes de substituição. A obra foi editada pela National Biomedical Research Foundation dos EUA e organizada por Dayhoff e colaboradores de 1965 a 1978, constituindo um marco na Bioinformática pela inestimável contribuição feita na comparação de seqüências pelo desenvolvimento de programas computacionais para detecção de seqüências divergentes e inferência de relações evolutivas, etc.

• 1966 – Marshall Nirenberg, Heinrich Mathaei e Severo Ochoa – O código genético foi decifrado pela demonstração de que uma seqüência de três bases de nucleotídeos, ou

seja um códon, determina cada um dos 20 aminoácidos, significando que há 64 combinações possíveis para os 20 aminoácidos.

• 1967 – WM Fitch e Margoliash – As primeiras árvores evolutivas a partir de seqüências de proteínas; estudo que foi chamado de Filogenética.

• 1968 – ARPA – Ao primeiros protocolos de rede Packet-switching network são apresentados.

• 1969 – ARPA – A ARPANET (o embrião da internet) é criada pela conexão de computadores em Stanford, UCSB, University of Utah e UCLA.

• 1970 – Needleman e Wunsch – Publicação dos detalhes do algoritmo de Needleman-Wunsch para comparação entre seqüências biológicas.

• 1971 – Ray Tomlinson (BBN) inventa o programa para e-mail.

• 1972 – Paul Berg e colaboradores na Universidade de Stanford (EUA) criam a primeira molécula de DNA recombinante pela combinação do DNA de dois organismos diferentes.

• 1973 – Herbert Boyer e Stanely Cohen desenvolvem a clonagem de DNA. Nesse mesmo ano, o Brookhaven Protein Data Bank (PDB) é anunciado (Acta. Cryst. B, 1973, 29: 1746) e Robert Metcalfe defende tese de doutorado (Ph.D.) na Universidade de Harvard (EUA) com a descrição da Ethernet.

• 1974 – Frederick Sanger desenvolve a técnica de sequenciamento de DNA. No mesmo ano, Vint Cerf e Robert Kahn desenvolvem o conceito de redes de conexão de computadores no que chamariam de “internet” com a criação do *Transmission Control Protocol* (TCP), enquanto Charles Goldfarb cria a SGML (*Standardized General Markup Language*). Também em 1974 é lancado o periódico **Nucleic Acids Research** (ISSN 1362-4962), um dos mais importantes veículos de comunicação nas áreas de genômica e bioinformática.

• 1975 – Bill Gates e Paul Allen fundam a Microsoft Corporation, enquanto P. H. O’Farrell (J. Biol. Chem., 250: 4007-4021, 1975) anuncia a eletroforese bi-dimensional em gel SDS poliacrilamida e E. M. Southern publica os detalhes experimentais da técnica de Southern Blot (J. Mol. Biol., 98: 503-517, 1975).

• 1976 – O *Unix-To-Unix Copy Protocol* (UUCP) é desenvolvido na Bell Labs.

• 1977 – A descrição completa do Brookhaven PDB (http://www.pdb.bnl.gov) é

publicada (Bernstein et al., J. Mol. Biol., 1977, 112:, 535). Allan Maxam e Walter Gilbert (Harvard) e Frederick Sanger (U.K. Medical Research Council) relatam métodos para sequenciamento de DNA. No mesmo ano, R. Staden publica um pacote de *softwares* para análise de seqüências.

• 1978 – David Botstein desenvolve a técnica de RFLP (*restriction fragment length polymorphisms*). No mesmo ano, o termo "***bioinformatics***" era cunhado pot Paulien Hogeweg para descrever o estudo dos processos de informação em sistemas bióticos. A **Bioinformatica** hoje compreende a criação e manutenção de bancos de dados, algoritmos, técnicas computacionais e estatísticas, além de teorias para solucionar problemas formais e práticos que surgem do gerenciamento e análise dos dados biológicos.

• 1980 – O European Molecular Biology Laboratory (EMBL) é estabelecido como biblioteca de dados para coletar, organizar e distribuir seqüências de nucleotídeos e dadfos correlatos. Tal função atualmente é realizada pelo European Bioinformatics Institute (EBI), Hinxton, U.K. Nesse mesmo ano, o primeiro gene de um organismo tem sua seqüência completa divulgada: o gene FX174 que consiste em 5,386 pares de bases codificando nove proteínas. Wüthrich e colaboradores publicam artigo detalhando o uso da Ressonância Magnética Nuclear (NMR) multi-dimensional para determinação da estrutura de proteínas (Kumar et al., Biochem. Biophys. Res. Comm., 1980, 95:, 1).

• 1981 – Desenvolvimento do algoritmo de Smith-Waterman; a IBM introduz o seu computador pessoal (*Personal Computer*, PC) no mercado. O conceito de motivo em uma sequência é apresentado (Doolittle).

• 1982 – O Genetics Computer Group (GCG) é criado como parte do Centro de Biotecnologia da University of Wisconsin. Ocorre a publicação da terceira versão do GenBank e é divulgado o seqüenciamento do genoma do fago lambda.

• 1983 – Lançado o algoritmo de busca de seqüências em bancos de dados – Wilbur-Lipman. No mesmo ano, o *Compact Disk* (CD) é lançado.

• 1984 – Jon Postel anuncia o *Domain Name System* (DNS) para postagem online. No mesmo ano, o computador Macintosh é anunciado pela Apple Computers.

• 1985 – Anunciado o algoritmo para comparação rápida de seqüências – FASTP/FASTN. No mesmo ano, Kary Mullis e colaboradores na Cetus Corporation publicam artigo sobre a polymerase chain reaction (PCR), criada dois anos antes.

• 1986 – Anunciado o banco de dados SWISS-PROT, criado pelo Departmento de Bioquímica Médica da University of Geneva e pelo European Molecular Biology Laboratory (EMBL).

• 1987 – Larry Wall cria a linguagem PERL (*Practical Extraction Report Language*), enquanto o primeiro "Yeast artificial chromosome" (YAC) é usado. Os YACs permitem a utilização da maquinária de duplicação de DNA das células. Neste mesmo ano, o termo **genômica** foi cunhado (como um nome para um novo periódico científico) por T.H. Roderick significando mapeamento e sequenciamento para analisar a estrutura e organização de genes em genomas. O mapa físico da *E. coli* é publicado (Y. Kohara, et al., Cell 51: 319-337).

• 1988 – Criação do ***National Center for Biotechnology Information*** (NCBI). O algoritmo FASTA para comparação de sequências é publicado por Pearson e Lipman. Um novo programa (um vírus de computador para a Internet desenvolvido por um estudante) infecta 6,000 computadores do exército dos EUA. A rede EMBnet para distribuição de bancos de dados é criada, enquanto o Genetics Computer Group (GCG) se torna uma empresa privada.

• 1990 – O método mais rápido e popular de alinhamento pareado de seqüências – BLAST – é lançado. A especificação HTTP 1.0 é publicada. Tim Berners-Lee divulga o primeiro documento HTML. O projeto Genoma Humano é lançado.

• 1991 – A estratégia de sequenciamento ESTs (Etiquetas de Seqüências Expressas, ou seja transcritas) é descrita por J. Craig Venter et. al., Science, 252: 1651-1656. O banco de dados Gene Ontology (GO) é criado. O instituto de pesquisas CERN em Genebra anuncia a criação dos protocolos que fazem a World Wide Web (www). **Linus Torvalds anuncia o sistema operacional baseado em Unix (Unix-Like) que depois veio a ser chamado Linux**[1]. Trabalhando em conjunto com o ganhador do prêmio Nobel Hamilton Smith, Craig Venter e outros criam o método '*shotgunning*'.

[1] Linus Torvalds foi um grande defensor do movimento 'open source', uma abordagem democrática para o desenvolvimento de softwares. Irreverente, o criador do Linux chegou uma vez a proclamar: "*Software is like sex; it's better when it's free*!"

• 1992 – Mel Simon e colaboradores anunciam o uso dos BACs (Bacterial Artifial *Chromosomes*) para clonagem. O Institute for Genomic Research (TIGR) é fundado por Craig Venter e publica o rascunho do genoma completo do *Haemophilus influenzae* (bactéria com quase 2 milhões de nucleotídeos) usando o recém divulgado método "shotgun".

• 1993 – Criação do Sanger Center, Hinxton, UK. Realização da primeira conferência *International Conference on Intelligent Systems for Molecular Biology* (ISMB) in Bethesda, Maryland.

• 1994 – Criação do EMBL European Bioinformatics Institute (EBI). A *Netscape Communications Corporation* é fundada e lança o Navigator, a versão comercial do Mozilla, o browser do NCSA. O banco de dados PRINTS (contendo motivos de proteínas) é publicado por Attwood e Beck.

• 1995 – Divulgação do seqüenciamento completo dos primeiros genomas bacterianos, incluindo *Haemophilus influenzae*. No mesmo ano, a Microsoft lança a versão 1.0 do Internet Explorer, a Sun lança a versão 1.0 do Java, enquanto a Sun e a Netscape lançam juntas a versão 1.0 do JavaScript. A versão 1.0 do Apache é liberada.

• 1996 – A ovelha clonada "Dolly" é anunciada. O genoma da levedura *Saccharomyces cerevisiae* (12.1 Mb) é sequenciado. O banco de dados Prosite é anunciado por Bairoch e colaboradores. A empresa Affymetrix produz os primeiros chips de DNA comerciais. A versão-rascunho da XML é lançada pela W3C.

• 1997 – Desenvolvimento do PSI-BLAST. Criação oficial da ISCB (*International Society for Computational Biology*) como uma derivação institucionalizada da conferência anual *International Conference on Intelligent Systems for Molecular Biology* (ISMB) que acontece desde 1993.

• 1998 – Craig Venter deixa o TIGR e junta-se a PE Corporation para criar a empresa Celera, em Rockville, Maryland, que anuncia o sequenciamento do genoma humano para daí a três anos. Isto precipita o trabalho do projeto público Genoma Humano. A Celera usa o seqüenciador Applied Biosystem's ABI Prism 3700 (uma máquina cinco vezes mais rápida e automatizada que as concorrentes da época). A FAPESP financia o sequenciamento do genoma completo da bactéria *Xylella fastidiosa*. O Swiss Institute of Bioinformatics é fundado. Lançamento do periódico **Journal of Bioinformatics** (01

January 1998 - 31 December 2001), que passou a se chamar **Current Affairs in Bioinformatics** (01 January 2002 - 31 December 2005) e atualmente **Bioinformatics** (ISSN 1460-2059), o principal na área de bioinformática e biologia computacional.

• 1999/2000 – Genomas completos da levedura *Saccharomyces cereviseae*, do *Caenorhabditis elegans* e da mosca *D. Melanogaster* (180Mb) foram os primeiros eucariotos anunciados.

• 2000 – Genomas completos da planta *Arabidopsis thaliana* (100 Mb) e da bactéria *Pseudomonas aeruginosa* (6.3 Mbp) são publicados. No Brasil é criada a **Rede Genoma Nacional (BRGene),** uma iniciativa do CNPq/MCT, compreendendo 25 laboratórios distribuídos em todo o país e coordenada por Andrew Simpson, do Instituto Ludwig de Pesquisas do Câncer em São Paulo.

• 2001 – Publicação do rascunho do Genoma Humano (*Homo sapiens,* (3,000 Mbp)). O projeto oficialmente começou em Outubro 1, 1990. No Nordeste é criado o **PROGENE (Programa Genoma Nordeste),** com financiamento do CNPq/MCT e FAPs da região, sendo coordenado por Paulo Andrade da Universidade Federal de Pernambuco.

• 2002 – João Meidanis e colaboradores fundam a empresa Scylla Bioinformatics em São Paulo, através da iniciativa de cinco bioinformatas oriundos do grupo que criou as soluções inovadoras que ajudaram a construir o sucesso dos primeiros projetos genoma brasileiros, como os das bactérias *Xylella fastidiosa* e *Xanthomonas citri*, do projeto EST da Cana-de-Açúcar, entre outros, permitindo ao Brasil ocupar posição de destaque na pesquisa genômica mundial.

• 2003 – Publicação do genoma completo da *Chromobacterium violaceum*, pela Rede Genoma Nacional na importante revista PNAS (*Proc Natl Acad Sci U S A 100: 11660–5*).

• 2004 – O genoma completo do *Rattus norvegicus* é publicado na edição de 1o. de Abril da revista Nature.

• 2005 – O genoma completo do chimpanzé *Pan trogloditis* é anunciado. Criação do periódico “oficial” da ISCB, com o advento da publicação no formato *open access*, em parceria com a *Public Library of Science*, o **PLoS Computational Biology**. Criação da Associação Brasileira de Bioinformática e Biologia Computacional (AB3C) e do primeiro evento oficial, o encontro anual X Meeting, em Caxambu, MG.

• 2006 – Tem início a nova era do *Next-generation DNA sequencing*, com a premiere do sequenciamento "flow cell" sendo o GS20 (454 Life Sciences), aparelho no qual uma única corrida forneceu dados de sequenciamento shotgun para montagem de-novo do genoma do *Mycoplasma genitalium* com 96% de cobertura em 99.96% de acurácia (Margulies et al. 2005). A **14ª Conferência Anual ISMB** da ISCB (**o maior evento mundial da Bioinformática**) **ocorre no Brasil**, no Centro de Convenções do Ceará, em Fortaleza, com quase mil participantes de todo o mundo, sendo pouco mais de cem destes brasileiros.

• 2007 – Terceiro genoma completo de um primata, o macaco rhesus é divulgado.

• 2008 – Publicação do genoma completo de um indivíduo humano por *massively parallel DNA sequencing* (Nature. 2008 Apr 17;452(7189):872-6.)

• 2009 – Surgimento dos sistemas de sequenciamento de DNA *'next-next' generation* tais como o *true single-molecule sequencing* ou tSMS (uma abordagem de sequenciamento-por-síntese em moléculas individuais), a abordagem baseada na *fluorescence resonance energy transfer* (FRET) ou o sequenciamento por nanoporos (Ver Lista de Empresas no Mercado Next-Generation – Apêndice IV e Tabela 1 abaixo). O website do NCBI lança o *Short Read Archive* (SRA) para armazenar dados brutos (*raw sequencing data*) oriundos das novas plataformas de sequenciamento "next generation" incluindo Roche 454 GS System®, Illumina Genome Analyzer® (ex-Solexa), Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, e outros.

Tabela 1 – Comparação de Métodos publicado por Holt e Jones, 2008.

Table 1. Manufacturer's specifications for instrument configuration and production of single end sequences from a single flow cell

Platform	Method	Template prep	Starting DNA (μg)	Instrument configuration	Throughput statistic	Data per run (Gbp)	Reagent cost per run ($)[a]	Run time
454 GS-FLX	Pyrosequencing	Emulsion PCR	3–5	Single picotiter plate, partitionable into 8 lanes	238-bp read[b]	0.1	8500	7.5h
Illumina 1G	Four-color SBS with reversible terminators[c]	Bridge PCR	0.1–1	Single flow cell, partitionable into 8 lanes	35-bp read	1.3	3000	3 d
ABI SOLiD	Oligonucleotide ligation with two-base, four-color encoding	Emulsion PCR	0.1–20	Independently controlled dual-flow cells, each partitionable into 8 lanes	35-bp reads, mapped to reference sequence allowing up to three mismatches	4	3400	7 d
Helicos HeliScope	Single-color SBS with virtual terminators	Not applicable	Not available	Single 25-lane flow cell	30-bp read	7.5	18,000	14 d

[a]Reagent costs are list prices.
[b]Average read length for a typical whole-genome library, using long read kit.
[c](SBS) Sequencing by synthesis.

1.2.2 A Natureza da Bioinformática

A Bioinformática combina a Matemática, a Computação e as Engenharias na exploração e compreensão dos dados biológicos gerados em larga-escala, tais como em experimentos de sequenciamento de genomas, expressão gênica (chips, SAGE, microarrays, etc) e proteomica (espectrometria de massas). Como uma disciplina em franca expansão, a Bioinformática se presta como o "braço exato" dos projetos genoma e pós-genomas que continuam produzindo imensas quantidades de dados biológicos a partir de diversos organismos, e, destarte, necessitam de armazenamento e tratamento desses dados em BDs públicos cada vez mais complexos e volumosos. Isto gera uma demanda para *datawarehouses* e sistemas de informação (SI) que acompanhem o crescimento exponencial desses BDs, ao lado da literatura cientifica correlata que também se avoluma acerca destes dados depositados. Assim, a Bioinformática tem tudo a ver com a idealização, construção e manutenção dessas ferramentas que visam, sobretudo, o incremento da capacidade analítica dos dados biologicos. Considerando a voracidade (velocidade e volume) com que novas seqüências são depositadas nos BDs de DNA e de proteínas a cada dia, existe uma evidente demanda pela conversão de toda essa informação em conhecimento real (genético, bioquímico ou biofísico). Essa conversão se dá ao se decifrarem as pistas (implícitas nos dados) estruturais, funcionais e evolutivas codificadas na linguagem das seqüências biológicas. O fluxo da informação genética (do DNA ao RNA e, deste, à proteína) se alinha perfeitamente ao fluxo da mineração de dados na Ciência da Computação (**dos dados** à **informação e destas ao conhecimento**).

As **seqüências**, p.ex., são os **dados**; a **informação** é o **resultado da análise** feita com aqueles dados, onde se buscam os significados; **o conhecimento é a interpretação ou extrapolação das informações**, consideradas num determinado contexo. Os dados, portanto, possuem uma dimensão atômica (ou individual), enquanto as informações assumem uma dimensão sintática (já que possuem um significado *per si*) e o conhecimento tem uma dimensão semântica, uma vez que representa uma reunião de significados dentro de um contexto (Figura 1.1).

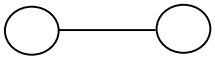
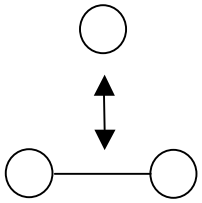
	DIMENSÃO	REALIDADE	OPERAÇÃO	EXEMPLO
Dados	**Atômica** 	**Observação, transformação**	Selecionar, Salvar, Recuperar, Visualizar	**Genes hiperregulados**
Informação	**Sintática** 	**Anotação, comparação**	Classificar, correlacionar	**Regiões genômicas regulatórias**
Conhecimento	**Semântica** 	**Interpolação, generalização**	Avaliar significância, Reduzir e minerar dados	**Padrões de ligação de TF**

Figura 1.1 – **Dado:** A representação de uma observação de um mundo natural em uma abstração computacional. **Informação:** A correlação de entidades de dados que compartilham uma característica. **Conhecimento:** A inferência dos aspectos de conceitos semânticos, quer dizer, modelos abstratos de domínio do problema, a partir de entidades de informação (Gentilmente cedido por Steipe, 2002;).

Dentro deste contexto, pode-se resumir a metodologia geral mais apropriada para os estudos em bioinformática como partindo dos seguintes passos:

• A formulação de hipótese(s) – testáveis *a posteriori*;
• O desenvolvimento de modelo(s) incorporando conhecimento prévio;
• A realização das análises;
• A "validação" dos resultados;
• O refinamento da(s) hipótese(s) / modelo(s).

1.2.3 Um Modelo Formalizado de Bioinformática

Sabe-se que um gargalo da pesquisa pós-genômica é a extração de informações a partir dos dados brutos, transformando-as em conhecimento aplicável. Um dos problemas para se atingir esse objetivo surge do desafio conceitual de se **analisar dados biológicos complexos em larga escala**. Os métodos computacionais precisam ser desenvolvidos e empregados de maneira que sejam relevantes para a

biologia e possam guiar a compreensão, de maneira confiável, sobre os contextos moleculares e celulares envolvidos. Os pesquisadores da área de biociências sempre esperaram que os padrões mais relevantes se tornariam óbvios quando se dispusesse de grande volume de dados, enquanto os cientistas da computação esperam que o ambiente biológico defina os problemas em termos computacionais. Ambas as expectativas carecem de maior fundamentação e representam o hiato existente entre as ciências biológicas e computacionais, o qual tem sido freqüentemente subestimado.

A Bioinformática não será produtiva como um simples processo de busca em BDs e coleta/catalogação de anotações a partir de uma infinidade de web sites, como parece ser o processo mais convencional atualmente em muitos laboratórios biológicos. Será preciso antecipar novas abordagens antes que elas se tornem óbvias, construindo tecnologias que apoiarão estas abordagens e garantirão trajetórias adequadas rumo ao objetivo final das abordagens. Este requerimento no campo da Bioinformática é altamente dinâmico e direcionado à inovação.

Esquematicamente, o **processo científico em Bioinformática** a ser utilizado pode ser resumido como o seguinte:

- Aquisição de dados
- Correlação com outros dados
- Conclusões extraídas a partir de relações quantificáveis
- Generalização
- Predição
- Validação (testes e verificações)
- Interpretação
- Inferência

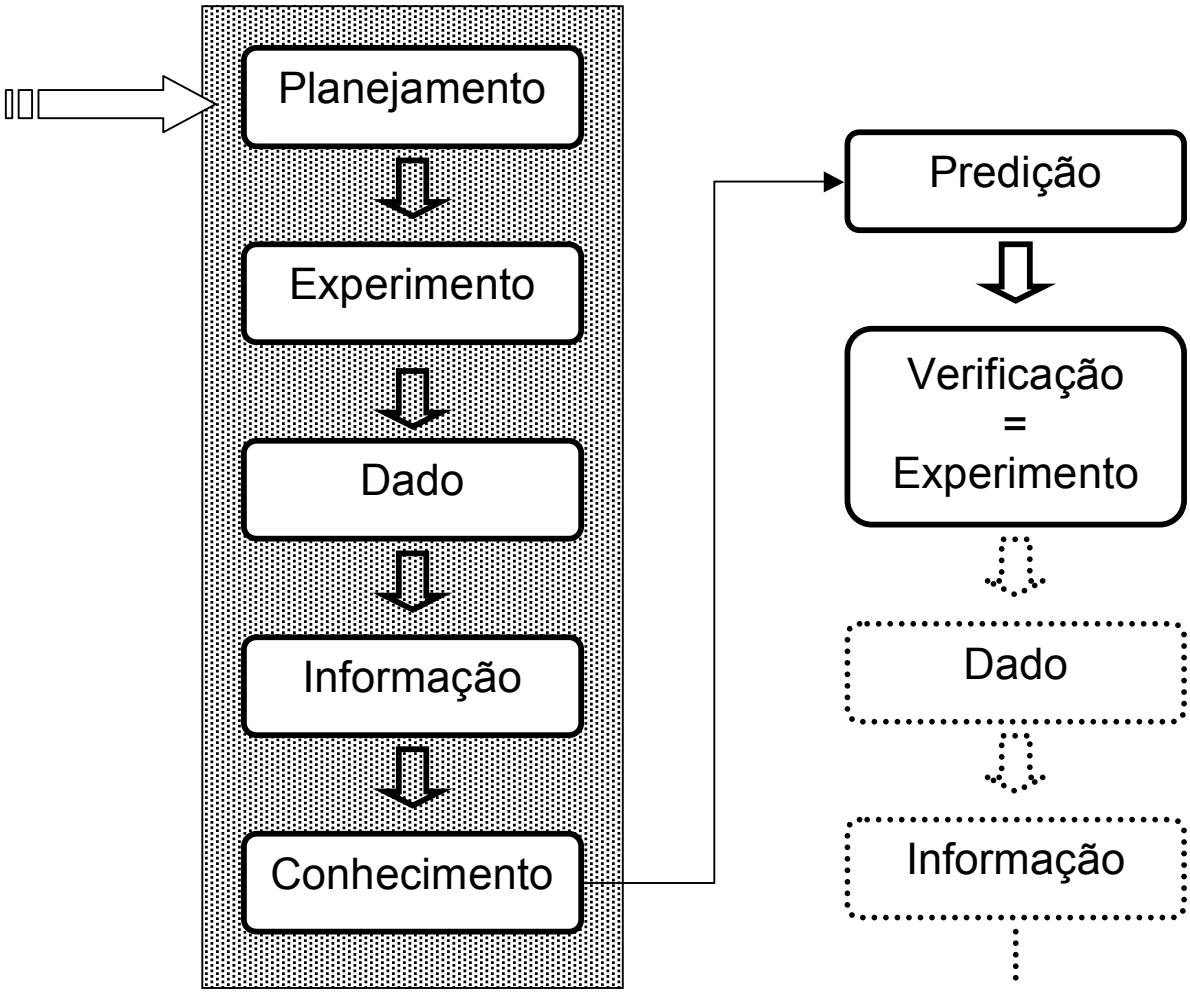


Figura 1.2 – Esquema ilustrativo do processo científico na Bioinformática, cuja aquisição de conhecimento final passa, inevitavelmente, por etapas anteriores (de tratamento dos dados, suas correlações e generalizações, extrapolações) como pressupostos para a interpretação viável e as possíveis hipóteses, formadas a partir de tais analises, que serão testadas na retro- alimentação do processo e novamente na validação das informações obtidas.

1.3. ÂMBITO DAS CIÊNCIAS GENÔMICAS E PÓS-GENÔMICAS

A genômica atualmente compreende o sequenciamento de genomas, a determinação do conjunto completo de proteínas codificadas por um organismo, e o funcionamento dos genes e vias metabólicas (*pathways*) em um organismo. Assim, a genômica não somente lida com a determinação da informação genética presente num organismo, mas também com a compreensão dos mecanismos pelos quais essa

informação é usada pelo organismo.

A informação gerada na genômica é enorme, fazendo com que a interpretação e o gerenciamento dessa informação requeiram o uso de grande poder computacional, tanto em hardware como em softwares específicos. Toda a bioinformática está voltada para a aquisição, armazenamento, análise, e visualização da informação biológica. Os BDs, para armazenamento e análise de informação genômica, são atualmente ferramentas essenciais para os geneticistas e bioquímicos. A Proteômica compreende o estudo dos produtos gênicos codificados por um genoma, incluindo uma lista de quais genes são expressos, o momento de sua expressão, o tipo e a extensão de qualquer modificação pos-tradicional da proteína, a função da proteína codificada,e sua localização em vários compartimentos celulares.

A disciplina da genômica pode ser dividida em dois principais domínios:

(1) genômica estrutural e
(2) genômica funcional.

1.3.1 A Genômica Estrutural lida com a determinação das sequências completas dos genomas ou o conjunto completo de proteinas produzidas por um organismo. Ela tem progredido em passos:
(1) pela construção de mapas físicos e genéticos de alta resolução,
(2) pelo sequenciamento de genomas, e
(3) pela determinação do conjunto completo de proteinas produzidas por um organismo; muitas vezes isso também vai incluir a determinação da estrutura tridimensional das proteínas em questão.

1.3.2 A Genômica Funcional estuda o funcionamento dos genes e das vias metabólicas em que seus produtos (gênicos) atuam, ou seja, os padrões de expressão gênica. Ela pode ser definida como a determinação da função, em última instância, de todos os produtos gênicos (ou seja, as proteínas) codificados pelo genoma de um organismo. Isto inclui soluções para questões que vão desde como um gene é

expresso, até como o seu produto gênico está relacionado em sequência e estrutura a produtos de outros genes do mesmo organismo, ou ainda como este gene interage com os outros?

Tais questões podem ser respondidas pelo estudo de: (i) onde e quando certos genes são expressos (*expression profiling* ou perfil de expressão); (ii) funções de genes específicos pela mutação seletiva dos genes desejados; e (iii) interações que ocorrem entre proteínas e entre estas e outras moléculas. Estas têm sido as principais orientações e rumos das pesquisas que os geneticistas moleculares têm conduzido desde que a genômica se instalou no cenário da investigação biológica. Ao invés de se concentrar em apenas um gene a cada vez, a genômica funcional procura examinar os genes presentes no genoma de uma vez só. Sendo assim, as técnicas usadas na genômica funcional possibilitam análises em larga-escala (*high throughput*) que permitem um acúmulo de dados muito rápido e, por sua vez, exigem recursos automatizados e computacionais para sua execução.

Um aspecto bem importante da genômica funcional é a determinação da função de sequências anônimas e de genes específicos. A melhor maneira de se alcançar isso continua sendo pela clonagem do gene, mutação dele *in vitro* e reintrodução do gene mutado no organismo hospedeiro para daí analisar o seu efeito. O uso de estratégias de larga-escala (*high throughput*) têm sido usadas para mutação de muitos genes no genoma, coleção de cepas mutantes e para formação de bibliotecas mutantes pan-genômicas (*genome wide mutant libraries*). Tais bibliotecas já foram feitas em diversas espécies modelo, tais como bactérias, levedura, plantas, e mamíferos, o que pode ser referido também como genômica mutacional.

1.3.3 Benefícios dos Projetos Genoma

Os projetos genoma permitem (ou possibilitam) a determinação completa da informação genética presente nos genomas sequenciados. As relações entre os genes podem ser deduzidas com bastante confiabilidade, ao mesmo tempo em que muitos detalhes sobre a organização dos genomas e a evolução dos indivíduos podem ser revelados, além dos mecanismos envolvidos. Os projetos genoma, na época de seu início (anos 90), abriram muitas possibilidades de pesquisa futura que hoje já são

realidade como a própria genômica funcional. Uma das expectativas mais aguardadas era a de que as sequências de genomas possibilitassem a descoberta de várias interações moleculares que conduzem ao desenvolvimento normal de células, tecidos, órgãos, o que permitiria, inclusive, a melhor compreensão de mecanismos patogênicos em diversas condições de interesse. Tal expectativa vem sendo bem cumprida pela genômica, especialmente em aspectos quantitativos tais como as informações contidas nos SNPs (*single nucleotide polymorphisms*) que não cessam de se acumular em dados disponíveis muito úteis para revelar suscetibilidade e resistência a diversas enfermidades como certos tipos de câncer, p.ex. O conhecimento mais apurado da base genética das doenças humanas deverá facilitar o seu tratamento e cura. Pode fornecer também respostas para as diferentes maneiras de reação a uma mesma substância / droga, o que convencionalmente se chama Farmacogenômica. Outrossim, o conhecimento advindo dos projetos genoma em microorganismos tem aumentado a compreensão sobre a patogenicidade de alguns deles, propiciando melhores métodos de prevenção e tratamento das doenças causadas por estes patógenos.

Não se pode ignorar que uma grande variedade de métodos, recursos e técnicas laboratoriais (avançados e aprimoráveis) foram (e continuam sendo) desenvolvidos no bojo dos projetos genoma. Isto representa um grande benefício porque indica que a demanda oriunda dos projetos e da pesquisa genômica ajuda a alavancar o estado-da-arte da biologia molecular em geral, uma vez que os constantes aperfeiçoamentos metodológicos exigidos pela ciência genômica fazem avançar o arsenal laboratorial disponível pela comunidade científica mundial.

1.3.4 Seqüenciamento, Montagem e Análise de Genomas

Um genoma pode ser definido como todo o conjunto de genes e DNA extra-gênico de um organismo, ou RNA, no caso de algumas famílias de vírus. As primeiras técnicas para o estudo de genomas visavam obter informações gerais sobre a sua composição (tamanho aproximado, porcentagem de nucleotídeos, número de cromossomos, localização de genes, etc.) já que não existiam metodologias para a obtenção em larga-escala das seqüências dos mesmos. Essa realidade mudou quando

foi desenvolvido o método de seqüenciamento automático de DNA, por sua vez baseado no principio de clonagem e amplificação de DNA (desenvolvido por Mullis, 1985). Os métodos mais modernos são capazes de produzir seqüências de, no máximo, 1000 bases (1kb) aleatórias dentro do genoma, o que significa que não existe conhecimento de qual região a seqüência gerada deriva. Um genoma de um organismo pequeno (um procarioto como a bactéria *Escherichia coli*, p. ex.) possui aproximadamente 10^7 bases, enquanto outros maiores (como a ameba de vida livre *Amoeba dubia*) podem alcançar até mesmo 10^{12}! Percebe-se, assim, alguns problemas que surgem com esse fato: a necessidade da realização de diversas reações de seqüenciamento para a obtenção da seqüência completa do genoma; o uso de algum tipo de metodologia para ordenar as seqüências corretamente, de modo a ter-se uma noção dos cromossomos ou do genoma inteiro. Em alguns genomas (os de eucariotos superiores, como mamíferos) pode-se encontrar até 98% de seqüências não-codificantes (aquelas que não produzem uma proteína), o que gera a necessidade de metodologias mais refinadas para encontrar os genes dentro do genoma.

A ordenação correta das seqüências, num contexto que se assemelhe a distribuição e densidade de genes ao longo do genoma, é o processo conhecido como montagem (*assembly*) de genomas. A montagem é feita através do uso de algoritmos[2] que irão alinhar as seqüências, sobrepondo redundâncias em suas extremidades para identificar as regiões comuns e agrupá-las (ou clusterizá-las) em contigs. Contigs ideais seriam, p.ex., os que representassem os cromossomos. Um exemplo de montagem de um pequeno trecho no genoma pode ser visto na Figura 1.2

Um terceiro tipo de gargalo existente seria a localização exata dos genes dentro do genoma (especialmente de eucariotos superiores), o que é atualmente resolvido pelo uso de metodologias gerais e complementares, tais como: a) uma estratégia de sequenciamento parcial em que o mRNA total é extraído, uma biblioteca de DNA complementar (cDNA) é construída e os transcritos são seqüenciados (como ESTs, p. ex.) para depois serem localizados nos genomas através de técnicas de alinhamento;

[2] Um algoritmo é um processo sistemático para a resolução de um problema; ele computa uma *saída*, o resultado do problema, a partir de uma *entrada*, as informações inicialmente conhecidas e que permitem encontrar a solução do problema. Durante o processo de computação, o algoritmo manipula *dados*, gerados a partir de sua entrada (Szwarcfiter e Markenzon, 1994)

b) características conservadas entre todos os genes nos genomas, tais como conteúdo GC, presença de promotores, presença de ORFs, etc., são utilizadas para identificar genes no genoma, numa evidente contribuição das áreas computacionais de reconhecimento de padrões e data mining.

Figura 1.3 – **Exemplo simplificado de montagem de seqüências.** A seqüência original de um gene (5' – 3') é mostrada acima, e os seis trechos de seqüências desconhecidas, obtidas através do seqüenciamento, podem ser vistas abaixo. O alinhamento das seqüências identifica áreas de sobreposição que são usadas para clusterizar/agrupar os resíduos na maior extensão contígua possível. Note que, neste exemplo ilustrativo, nenhum trecho da seqüência original deixou de ser coberto pelas seqüências menores, o que, nem sempre, ocorre na rotina dos trabalhos de clusterizacao.

Sequências e Nomenclatura – Um dos maiores desafios e dificuldades da bioinformática é a apresentação de grandes volumes de dados em um formato eficiente e facilmente compreensível. As sequências de nucleotídeos e de aminoácidos são facilmente reduzíveis a dados digitais pelo uso de códigos de letra única (*single letter codes*). O sistema de nomenclatura adotado na bioinformática se baseia nas recomendações da *International Union of Pure and Applied Chemistry* (IUPAC).

1.3.5 Anotação Gênica Após o Sequenciamento de Genomas

O objetivo principal e definitivo de todos os esforços empregados em seqüenciamento é descobrir funções moleculares (genéticas e bioquímicas) e celulares de todos os produtos gênicos codificados por estas seqüências. A interpretação da informação contida nas seqüências, isto é, a anotação gênica, é entretanto, uma tarefa não trivial e tem sido objeto de intensa pesquisa. *A priori*, a anotação gênica pode ser dividida em três etapas: a anotação no nível de nucleotídeo, a anotação no nível protéico e a anotação no nível de processos. A fase inicial da anotação, feita no nível de nucleotídeos tem como atividade principal a localização de marcadores através de

mapeamento e a procura de genes na seqüência de DNA. Nesta fase são primeiro identificados marcadores produzidos através de mapeamentos feitos por análises genéticas, citogenéticas ou de híbridos de radiação. Este conjunto de marcadores funciona então como pontos de referência para a análise subseqüente: a procura por genes. Uma vez identificados os genes, são então identificadas seqüências correspondentes a RNAs não codificantes, seqüências regulatórias, elementos repetitivos e polimorfismos. Após a anotação no nível de nucleotídeos, inicia-se a etapa de anotação no nível protéico. Esta etapa é constituída da nomeação das proteínas do organismo e associação de possíveis funções a estas proteínas. Neste caso, são utilizados bancos de dados de seqüências primárias, estruturais, de famílias gênicas ou de domínios funcionais como as bases SWISS-PROT, *Protein Data Bank* (PDB) ou PFAM.

Depois destes dois níveis tem início então a etapa de anotação no nível de processos. Esta etapa tem como objetivo relacionar o genoma a processos biológicos, isto é, estabelecer como os constituintes de um genoma se relacionam com o ciclo celular, a morte celular, embriogênese, metabolismo e manutenção da saúde do organismo. Este processo depende da existência de um BDs dotado de um esquema de classificação associado a funções biológicas conhecidamente descritas, com especificidade suficiente para distinguir entre proteínas que sejam membros de uma mesma família gênica. O BDs Gene Ontology (GO), criado em 1991, é um repositório desta natureza. A procura por genes codificantes de proteínas tem sido amplamente utilizada por vários projetos de genômica funcional. Esta etapa é, em geral, realizada em genomas de procariotos sem maiores dificuldades, uma vez que ela consiste basicamente na identificação de janelas/matrizes abertas de leitura (ORFs) na seqüência produzida. Em eucariotos, por outro lado, o processo de busca de genes é complicado pela presença de íntrons e sítios de *splicing* alternativo. Por essa razão, métodos diversos para a predição de genes em seqüências eucarióticas têm sido amplamente utilizados, de maneira geral, a procura por genes é feita a partir de dois métodos de predição distintos designados respectivamente, extrínsecos e intrínsecos.

Em genomas recém seqüenciados, genes são anotados primariamente com base em sua homologia com proteínas já caracterizadas em outros genomas. Este

enfoque é designado como extrínseco por desconsiderar as características existentes na seqüência investigada. Os programas baseados em busca de homologia, que são utilizados neste tipo de abordagem, têm como premissa a conservação existente entre as seqüências de diferentes espécies. Tais programas utilizam sensores que exploram a similaridade existente entre uma região genômica desconhecida e uma seqüência de proteína ou nucleotídeos presente em um BDs, para determinar se a região em questão é ou não uma região codificadora. Para detectar a similaridade entre seqüências, estas são alinhadas em um processo que consiste na comparação de duas seqüências diferentes do mesmo organismo, ou de organismos diferentes, para gerar um alinhamento local ótimo. Alinhar duas seqüências consiste em estabelecer uma correspondência entre as bases dessas seqüências de modo que a ordem não seja violada. Por ordem entende-se que as bases nas posições *n1* e *n2* (*n1* < *n2*) de uma seqüência estão associadas respectivamente às bases nas posições *m1* e *m2* da outra seqüência (*m1* < *m2*). Os algoritmos para predição intrínsecos são baseados em padrões de reconhecimento de características específicas do gene em associação com a análise do conteúdo da seqüência. Características específicas da seqüência normalmente associadas a presença de genes (promotores, códons inciadores e finalizadores, sítios de *splicing*, etc.) são utilizadas como sinais para inferir a presença de um gene juntamente com a distribuição de nucleotídeos que apresenta diferenças em regiões que contém genes e regiões intergênicas. A combinação da informação proveniente destes padrões permite não só a localização de genes completos em uma seqüência genômica como também de estruturas gênicas parciais nas extremidades da seqüência analisada.

1.4. APLICAÇÕES E ABORDAGENS DA BIOINFORMATICA

A Bioinformática às vezes pode ser referida como “Biologia Computacional”, mas vale a pena destacar que, de modo estrito, a Bioinformática seria mais como uma subdivisão ou um capitulo da Biologia Computacional. Esta última sendo, então, uma área muito maior e mais abrangente que integraria outros ramos da pesquisa e

desenvolvimento (P&D) em tópicos como tratamento de imagens biomédicas, bioengenharia, etc. A Bioinformática assume uma conotação mais precisa na questão da análise genômica e proteômica, não apenas na catalogação de componentes celulares em tecidos e organismos, mas também na compreensão da organizacao dos genomas, e de como as células, tecidos e organismos funcionam na saúde e nas doenças. Com o maciço volume de dados genômicos e pos-genômicos existente atualmente, não ha tempo/espaço para conversão de diferentes formatos para tantos dados. Isto exige padrões (rígidos, porém acessíveis) para a integração de dados e para sua disponibilizacao de forma gráfica amigável ao usuário das áreas biológicas. A Bioinformática é essa área, cuja P&D irá permitir a combinação eficiente das Ciências Biológicas e da Computação no intuito de dar sentido a essa avalanche de dados que os pesquisadores em Biotecnologia precisam explorar!

O impacto dos projetos genoma nos últimos anos não se resume a um simples aumento estrondoso de dados (as seqüências acumuladas), mas se estende a uma diversificação no tipo de dados moleculares. Uma seqüência completa de genoma apresenta não somente o conjunto inteiro dos genes e suas localizações precisas no(s) cromossomo(s), mas também as relações de similaridade de genes dentro daquele genoma e em comparação com outras espécies. O sequenciamento automático de DNA teve um impacto surpreendente na fronteira da geração de dados em larga-escala — *expressed-sequence tags* (ESTs), *microarrays* e *single-nucleotide polymorphisms* (SNPs), p.ex. among others. Até o momento, a bioinformática tem sido uma disciplina pragmática através da qual se expressam as demandas por tecnologias da informação e da computação na produção de dados em larga-escala na genômica e pós-genômica. Todavia, à medida em que os dados são convertidos a conhecimento e as regras empíricas levam aos princípios, a bioinformática se prepara para tornar-se uma disciiplina mais fundamental. Provavelmente, a bioinformática compreenderá aspectos não apenas biológicos e práticos da informática (Ciência da Computação), mas também e principalmente fundamentos teóricos da matemática para ajudar na detecção de arquiteturas básicas dos sistemas complexos de informação biológica, alem da integração dos princípios físicos e químicos aos princípios biológicos. Quando for

possível ter representações computacionais completas de células e organismos vivos e saber todos os princípios de como eles “computam”, então estaremos na época em que, conforme Sydney Brenner previa, "a biologia computacional haverá se tornado computação biológica".

Referências

[1] Minoru Kanehisa1 & Peer Bork. Bioinformatics in the post-sequence era. Nature Genetics 33, 305 - 310 (2003). doi:10.1038/ng1109

[2] Guilherme P. Telles, Marília D. V. Braga, Zanoni Dias, Lin T. Li, José A. A. Quitzau, Felipe R. da Silva, e João Meidanis. Bioinformatics of the Sugarcane EST Project. Genetics and Molecular Biology, 24(1-4):9-15, 2001.

[3] Andrew J. G. Simpson et al. The genome sequence of the plant pathogen Xylella fastidiosa. Nature, 406:151-157, 2000.

[4] Ana C. R. da Silva et al. Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. Nature 417, 459-463 (23 May 2002).

[5] Christian Baudet e Zanoni Dias. Analysis of slipped sequences in ESTs Projects. Genetics and molecular research, v. 5, n. 1, p. 169-181, 2006.

[6] Minoru Kanehisa1 & Peer Bork. Bioinformatics in the post-sequence era. Nature Genetics 33, 305 - 310 (2003). doi:10.1038/ng1109.

[7] Suresh Kumar (2009). Bioinformatics web. Retrieved March 14, 2009 from: *http://www.geocities.com/bioinformaticsweb/*

[8] João C. Setubal e João Meidanis. Introduction to Computational Molecular Biology, Boston: PWS Publishing Company, 296pp., 1997.

[9] Miguel Galves e Zanoni Dias. Comparison of Genomic DNA to cDNA Alignment Methods. In: Brazilian Symposium on Bioinformatics 2005, 2005, São Leopoldo - RS. Lecture Notes in Bioinformatics. Berlin - Alemanha : Springer-Verlag, 2005. v. 3594. p. 170-180.

[10] Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., Jaffe, D.B. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 18:763–770.

[11] *Baxevanis, A.D. and Ouellette, B.F.F., eds. 2001 . Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins. John Wiley and Sons, Inc., NY*

[12] *Brenner, S., Lewitter, F., Patterson, M., and Handel, M., eds. 1998 . Trends Guide to Bioinformatics. Elsevier Science*

[13] *Aravind, L. and Koonin, E.V. 1999 Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J. Mol. Biol., 287:1023-1040.*

[14] *Altschul S.F., and Koonin, E.V. 1998. Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. TIBS 23(11):444-7.*

[15] *Pickeral, O.K. and Boguski, M.S. 1999. Book Review: The Bioinformatics bookshelf: teach yourself computational biology? Cell 96:451-455.*

[16] *Achuthsankar S Nair Computational Biology & Bioinformatics - A gentle Overview, Communications of Computer Society of India, January 2007*

[17] *Aluru, Srinivas, ed. Handbook of Computational Molecular Biology. Chapman & Hall/Crc, 2006. ISBN 1584884061 (Chapman & Hall/Crc Computer and Information Science Series).*

[18] *Cristianini, N. and Hahn, M. Introduction to Computational Genomics, Cambridge University Press, 2006. (ISBN 9780521671910 | ISBN 0521671914)*

[19] *Gilbert, D. Bioinformatics software resources. Briefings in Bioinformatics, Briefings in Bioinformatics, 2004 5(3):300-304.*

[20] *Keedwell, E., Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems. Wiley, 2005. ISBN 0-470-02175-6*

[21] *Brazilian National Genome Project Consortium. (2003). "The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability.". Proc Natl Acad Sci U S A 100: 11660–5. doi:10.1073/pnas.1832124100*

[22] *Mardis ER. The impact of next-generation sequencing technology on genetics.Trends Genet. 2008 Mar;24(3):133-41. Epub 2008 Feb 11. Review.*

[23] *Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 2008 Feb 28;9:128.*

[24] *Droege M, Hill B. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol. 2008 Aug 31;136(1-2):3-10.*

[25] *Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008 Apr 17;452(7189):872-6.*

[26] Holt RA, Jones SJ. The new paradigm of flow cell sequencing. Genome Res. 2008 Jun;18(6):839-46. Review.

Apêndice IV

Sumário de Empresas Atuantes no Mercado de Sequenciamento *Next-Next Generation*

A lista abaixo ilustra o grande numero de empresas atuantes no mercado de produtos para sequenciamento de DNA de nova geração, em franca oposição ao mercado restrito que era exercido por poucas companhias quando do inicio dos seqüenciadores automáticos de DNA pelo método de Sanger.

Company	**Web address**
23andMe	http://www.23andme.com
454 Life Science	http://www.454.com
ACGT	http://www.acgtinc.com
Affymetrix	http://www.affymetrix.com
Agencourt Biosciences Corporation	http://www.agencourt.com
Agilent Technologies	http://www.agilent.com
Applied Biosystems	http://www.appliedbiosystems.com
Beckman Coulter	http://www.beckmancoulter.com
Biotage	http://www.biotagebio.com
Brukner Daltonics	http://www.bdal.com
Clontech	http://www.clontech.com
Cogenics	http://www.cogenics.com
DeCode Genetics	http://www.decode.com
DNADirect	http://www.dnadirect.com
DNAstar	http://www.dnastar.com
Expression Analysis	http://www.expressionanalysis.com
GE Healthcare	http://www.gehealthcare.com
Genomic Solutions	http://www.genomicsolutions.com

Geospiza http://www.geospiza.com
Helicos Biosciences http://www.helicosbio.com
Illumina http://www.illumina.com
Intelligent Bio-systems http://www.intelligentbiosystems.com
Invitrogen http://www.invitrogen.com
Licor Biosciences http://www.licor.com
LingVitae AS http://www.lingvitae.com
Maxim Biotech http://www.maximbio.com
Nabsys http://www.nabsys.com
Nanogen http://www.nanogen.com
New England Biolabs http://www.neb.com
Nimblegen Systems http://www.nimblegen.com
Nugenhttp://www.nugenic.com
Open Biosystems http://www.openbiosystems.com
Operon Biotechnologies http://www.operon.com
Pacific Biosciences http://www.pacificbiosciences.com
Perkin Elmer Life Sciences http://www.perkinelmer.com
Promega http://www.promega.com
Qiagen http://www.qiagen.com
Reveohttp://www.reveo.com
Roche Applied Sciences http://www.roche.com
Saturn Biotech http://www.saturnbiotech.com.au
Sequenom http://www.sequenom.com
SeqWright http://www.seqwright.com
Stratagene http://www.stratagene.com
Taconic http://taconic.transnetyx.com
TeleChem International http://www.arrayit.com
Transgenomic http://www.transgenomic.com
Visigen Biotechnologies http://www.visigenbio.com

Fonte: Nathan Blow. DNA sequencing: generation next-next. Nature Methods - 5, 267 - 274 (2008) doi:10.1038/nmeth0308-267