e-Science Bioinformatics

Luciano Milanesi
Institute of Biomedical Technology CNR

ABSTRACT. The Bioinformatics community was an early adopter of the Internet and contributed to the e-Science development by publishing its vast amount of data, and conveying the rich interconnectedness of biological information. Bioinformatics is contributing in many research fields like: Genomics, Proteomics, Transcriptomics and applications in Molecular Dynamics. The enormous potentiality of Bioinformatics and its potential impact on Molecular Medicine, Biology, Biotechnology and Industry are introduced. The main Bioinformatics Institutions and Internet Resource Database available are provided.

KEYWORDS: *Bioinformatics, Genomics, Proteomics, Human Genome, e-Science*

Introduzione

Con i recenti progressi nell'ambito biomedico è diventato essenziale assicurare un supporto adeguato alle ricerche nell'ambito della medicina e delle scienze della vita; infatti una caratteristica dell'era post-genomica dipenderà da nuove modalità di trattamento dell'enorme mole di dati generata quotidianamente al fine di correlare le informazioni genotipiche con quelle fenotipiche e le analisi clinico-mediche.
I termini *bioinformatica* e *biologia computazionale* sono spesso usati in modo intercambiabile; informatica medica è un campo del tutto a se stante. L'informatica medica tratta generalmente dati prodotti dall'acquisizione strumentale in ambito ospedaliero come ad esempio immagini mediche (Immagini radiologiche, TAC, NMR, PET, Risonanza Magnetica ecc.), informazioni provenienti da cartelle cliniche dei pazienti, fino al livello della popolazione, mentre la bioinformatica tende ad occuparsi di informazioni a livello cellulare e di strutture e sistemi biomolecolari. Anche se i tre termini: bioinformatica, biologia computazionale e infrastruttura bioinformatica sono spesso usati in modo intercambiabile, in generale, essi possono venir definiti come segue: la *bioinformatica* si riferisce ad attività connesse a database, che hanno a che fare con insiemi di dati persistenti che vengono mantenuti in stato coerente lungo periodo di tempo essenzialmente indefiniti.
La *biologia computazionale* comprende l'uso di strumenti algoritmici per facilitare l'analisi biologica dei sistemi complessi.
L'*infrastruttura bioinformatica* comprende l'intero insieme dei sistemi di gestione informativa, strumenti analitici e reti di comunicazione che

supportano la biologia. Dunque, quest'ultima può essere considerata come un'impalcatura computazionale delle precedenti.
La bioinformatica è diventata la disciplina di riferimento in supporto della genomica, proteomica ed indagini genetiche a scopo diagnostico. Il bioinformatico è quindi un esperto che non solo conosce come impiegare gli strumenti della bioinformatica, ma cha sa anche come scrivere interfacce per ottenere un uso efficace dei dati generati dai vari esperimenti. Un tecnico bioinformatico, viceversa, è un soggetto addestrato che limita la sua conoscenza all'impiego di strumenti per la bioinformatica senza averne una comprensione approfondita.
Nel panorama internazionale si è assistito ad un massiccio investimento nella Bioinformatica da parte di Stati Uniti, Canada e Giappone, mirato soprattutto alla creazione di centri, reali o virtuali, che forniscano servizi e conoscenze alla comunità mentre, allo stesso tempo, spingono in nuove direzioni grazie ad attività di ricerca avanzate.
L'Italia è ben collocata nel panorama europeo dal punto di vista scientifico, anche se le risorse limitate, rispetto soprattutto a Germania e Regno Unito, hanno finora reso difficile l'implementazione di progetti indipendenti a grande respiro e di risonanza internazionale.
Oggi si è creata un'enorme attesa, soprattutto negli Stati Uniti, in Canada, Giappone e Europa per un impatto economico crescente della bioinformatica. Per esempio, grazie all'individuazione di nuovi agenti terapeutici e farmacologici ci si aspetta che la bioinformatica sia in grado di aumentare il numero di bersagli terapeutici da circa 400 a circa 4000 nei prossimi vent'anni e quindi, già solo grazie a questo, ci si aspetta che i proventi dell'industria farmaceutica aumentino di un fattore mille. Stime conservative fanno assumere che la Bioinformatica, solo negli Stati Uniti, permetterà di generare almeno 7 miliardi di dollari nei prossimi tre anni e che il mercato crescerà del 20% entro il 2006.
Lo sviluppo di piccole e medie aziende con finalità bioinformatiche in Europa è, come ci si può aspettare, proporzionale all'investimento in ricerca dei vari Stati e quindi molto attivo in Inghilterra, Francia, Germania, in crescita in Spagna, Olanda e paesi nordici.
Questa è solo una della possibili applicazioni della bioinformatica, che ci si aspetta abbia un impatto anche sullo sviluppo di vaccini, nuove biotecnologie, nanotecnologie, ecc. La tendenza è che gli investimenti industriali in bioinformatica siano triplicati entro il 2010, prevalentemente nel settore del software di analisi e delle banche dati specializzate.
La biologia e la medicina hanno da sempre cercato di studiare l'organismo umano a tutti i livelli, sia morfologico sia funzionale, per il miglioramento della salute umana. Con l'introduzione delle più moderne tecnologie di biologia molecolare si sono iniziate a comprendere meglio le regole d'espressione genetica, i vari passaggi metabolici e la struttura del DNA, RNA e delle proteine.

Applicazione della Bioinformatica allo studio del Genoma Umano

L'enorme complesso d'informazioni, generato dal progetto internazionale sul sequenziamento del Genoma Umano, consentirà di comprendere il flusso delle informazioni che governano il passaggio dal Genoma al Fenotipo di un organismo.
Un utilizzo appropriato di questi dati, associato ad appositi programmi d'analisi, porterà nuove possibilità per la comprensione dell'espressione dei geni, della loro regolazione e delle malattie genetiche a loro correlate in caso di mutazioni o disfunzioni metaboliche. La disciplina che si occupa di queste problematiche per il trattamento dell'informazione biologica a tutti i livelli è la Bioinformatica.
La Bioinformatica si occupa, quindi, dell'acquisizione, memorizzazione, distribuzione, analisi e interpretazione dei dati, prevalentemente nell'ambito della biologia molecolare, con collegamenti sempre più importanti con la medicina. Questa nuova disciplina scientifica utilizza metodi di matematica, informatica, biologia, medicina, fisica allo scopo di migliorare la comprensione dei fenomeni biologici. Di seguito, si elencano, a titolo di esempio, alcuni dei principali obiettivi della Bioinformatica:

Sviluppo di strumenti per la generazione e il mantenimento dell'informazione proveniente dalle varie fonti:

- mappa fisica, mappa genetica, mappa cromosomica, mappa citogenetica, polimorfismi e informazione relativa alle sequenze genomiche e proteiche;
- raccolta e organizzazione delle informazioni genetiche associate alle patologie mediche;
- sviluppo di programmi di calcolo per l'analisi delle sequenze;
- sviluppo d'interfacce grafiche in grado di visualizzare in maniera efficace l'informazione richiesta;
- sviluppo di metodi software che consentano di agevolare tutte le fasi dei progetti;
- sviluppo di strutture per database specializzati ed integrati;
- sviluppo di strumenti informatici, includendo software e hardware e algoritmi per l'organizzazione e l'analisi dei dati;
- realizzazione di standard per lo scambio e la descrizione dei dati;
- realizzazione di una rete dati per la raccolta, la distribuzione e l'aggiornamento costante di tutta l'informazione prodotta;
- raccolta della bibliografia, brevetti e altri database di supporto all'informazione specifica;
- predizione dei geni nelle sequenze di DNA;
- predizione delle strutture tridimensionali delle proteine partendo dalle sequenze primarie;

- predizione delle funzioni biologiche e biofisiche sia dalle sequenze che dalle strutture;
- simulazione dei processi metabolici e cellulari basati su queste funzioni;
- realizzazione di sistemi per la correlazione dell'informazione in sistemi biologici complessi.

Da questo elenco risulta evidente che uno dei principali scopi della Bioinformatica è di fornire in tempi rapidi le informazioni e le metodologie d'indagine che consentano, ad esempio, di fornire alla scienza medica le necessarie informazioni per comprendere i meccanismi alla base di tutte le possibili disfunzioni d'origine genetica. Dalle mutazioni in regioni funzionali del DNA, alla mancata produzione di una certa proteina a causa di un anomalo funzionamento dei fattori di trascrizione, fino a comprendere il funzionamento d'ogni singolo gene in relazione con gli altri geni nei diversi processi metabolici che continuamente avvengono nel corso dell'intera vita. L'industria farmaceutica utilizza queste informazioni e metodologie e conoscenze al fine di produrre medicine, proteine specifiche e terapie geniche in grado di intervenire in maniera selettiva per risolvere le possibili cause di malattie.
Inoltre, la gran parte dei prodotti genici del genoma umano ha più di una funzione, in alcuni casi addirittura in antagonismo con altre. Per questo è necessario identificare i modelli quantitativi che stanno alla base di questi processi. In quest'ottica i metodi e i programmi messi a disposizione dalla Bioinformatica giocano un ruolo fondamentale nello sviluppo delle Biotecnologie Molecolari e della nuova Medicina Translazionale.

Sviluppo di una nuova disciplina scientifica

Per ottenere i precedenti obiettivi, il mondo della bioinformatica ha necessità di reperire personale giovane con un'adeguata formazione scientifica. Si riscontra, infatti, un gap in biologia e informatica; la creazione di un percorso di studi che istruisca in entrambi gli aspetti darebbe una risposta concreta alle esigenze del comparto industriale.
Le industrie operanti nel comparto delle biotecnologie hanno una forte dipendenza dalla ricerca e il settore della bioinformatica e dell'informatica medica non fanno eccezione. Un importante vantaggio competitivo sarebbe rappresentato dalla possibilità di incrementare in maniera sostanziale le collaborazioni sia a livello di ricerca di base che applicata.
Aree di interesse sono, ad esempio, la ricerca di nuovi sistemi ed algoritmi per l'identificazione funzionale di geni e proteine, l'adozione di tecniche e strumenti per l'estrazione automatica dell'informazione da basi di dati testuali quali ad esempio letteratura scientifica, classificazione automatica delle proteine, sviluppo di strumenti per la definizione/esecuzione di esperimenti "in silico" e la gestione automatica di flussi di lavoro soprattutto in collegamento a sistemi di calcolo distribuiti. Pertanto, l'industria ha necessità di luoghi dove imprese, laboratori

universitari, istituzioni economiche e di ricerca operino a stretto contatto al fine di:

- sviluppare ricerca fondamentale e a carattere applicativo;
- mettere a punto nuove tecnologie, prodotti e processi;
- valorizzare i risultati della ricerca e trasferire tecnologie e innovazioni al mondo produttivo;
- creare reti di cooperazione nazionali e internazionali;
- sviluppare attività economiche e imprenditoriali ad alta intensità di conoscenza.

In parole povere, occorrono centri di eccellenza capillarmente distribuiti sul territorio con una serie di servizi (cataloghi del know-how e delle tecnologie pronte all'uso, panel di esperti e consulenti, seminari e workshop, ecc.) per estendere e sviluppare i cosiddetti CLUSTER e Reti tematiche. Per mantenersi al livello presente non basta però più il lavoro di gruppi indipendenti: l'aumento enorme dei dati disponibili rende oggi essenziale la presenza di infrastrutture e, soprattutto, di risorse che oggi sono più carenti in Europa rispetto agli Stati Uniti e al Giappone. L'industria ha necessità di reperire personale giovane con un'adeguata formazione scientifica. Si riscontra, infatti, un gap tra biologia ed informatica; la creazione di un percorso di studi che cumuli entrambi gli aspetti darebbe una risposta concreta alle esigenze del comparto industriale.

Nell'ambito di una sanità "Informata e Formata" sulle specifiche ricerche di Bioinformatica è doveroso sensibilizzare i medici di medicina generale, i biologi e gli altri operatori interessati alla materia, su argomenti che possano essere applicati alla pratica clinica. Pertanto si mira a creare strumenti informatici che permettano alla comunità medico-scientifica e al mondo industriale di:

- analizzare e utilizzare i dati sui genomi per lo sviluppo di nuove terapie, medicinali e diagnosi sanitarie sempre più accurate;
- raccogliere le informazioni al fine di analizzare e utilizzare i dati sui genomi per lo sviluppo di nuove terapie, medicinali e diagnosi sanitarie sempre più accurate.

Sul piano della ricerca, è necessario lo sviluppo di sistemi software per la Bioinformatica, diretti ad esempio:

- alla modellizzazione della struttura tridimensionale a livello atomico delle molecole biologiche;
- all'analisi, lo studio e la predizione dei modi di interazione delle molecole biologiche;
- alla classificazione di dati basati su metodi di intelligenza artificiale quali le reti neurali, gli Hidden Markov Models, le Support Vector machines, ecc. mirati all'assegnazione di funzione o di caratteristiche funzionali ai prodotti genici di interesse;
- alla visualizzazione, analisi, confronto e ricerca in banche dati di

interazioni tra molecole biologiche e di pathway metabolici;
- alla correlazione dei dati di interazione con dati disponibili sulle basi genetiche di patologie complesse;
- all'identificazione di elementi regolativi della trascrizione all'interno e nell'intorno di geni;
- all'identificazione delle reti di regolazione genetica a partire da dati di espressione dalla conoscenza degli elementi regolativi. Questi algoritmi permetteranno di capire come i geni agiscono in maniera concertata e pertanto di ricostruire i pathway biologici in cui essi sono coinvolti e quindi capire la loro funzione;
- all'identificazione di siti di legame di fattori di trascrizione allo scopo di fornire un insieme di strumenti flessibili per prelevare i dati da sorgenti multiple;
- all'identificazione di reti di regolazione genetica. Una rete genetica può essere rappresentata matematicamente utilizzando equazioni differenziali ordinarie i cui parametri sono incognite che bisogna stimare utilizzando i dati sperimentali quali profili di espressione genetica per identificare le interazioni tra geni;
- all'interpretazione automatica di letteratura scientifica e di divulgazione delle più note riviste internazionali;
- alla modellistica biomedica e all'apprendimento automatico nella systems biology. A tal fine occorre impiegare metodologie innovative per la generazione di ipotesi su reti di regolazione genica e per lo studio e la simulazione di particolari processi cellulari;
- allo sviluppo di metodi basati su modelli qualitativi e metodi probabilistici, in grado di tener conto di informazioni incerte e conoscenza imperfetta sul dominio allo studio;
- allo sviluppo di sistemi informativi sanitari correlati a dati genetici. In questa attività verranno utilizzati ed adattati metodi ed approcci sviluppati nell'ambito dei sistemi informativi sanitari per tener conto delle esigenze specifiche delle applicazioni che trattano dati genetici, genomici e di proteomica;
- allo sviluppo di cartelle cliniche elettroniche orientate alla gestione ed all'integrazione dei dati clinici con dati genetici, di genomica funzionale e di proteomica;
- alla progettazione di sistemi informativi per la conduzione di studi di caratterizzazione genetico-clinica;
- alla gestione di dati e database orientati agli aspetti di privacy e sicurezza informatica;
- a costruire workflow per il design e lo sviluppo di applicazioni di supporto al lavoro collaborativo in bioinformatica;
- a sviluppare metodologie basate sul GRID computing e calcolo avanzato specifiche per la bioinformatica, per fare fronte a calcoli di notevole complessità algoritmica.

Risorse in internet utili in Bioinformatica

Una veloce panoramica dei termini e dei concetti base della genetica faciliterà la comprensione dei database di sequenze. Il sito della NCBI Genetics Review offre un'ottima panoramica dei concetti oltre a elencare alcuni ottimi riferimenti per informazioni addizionali. I termini seguenti sono essenziali per la comprensione della bioinformatica. Per un approfondimento di tematiche specifiche si può approfondire sui seguenti siti in Internet:

SCIENCE MAGAZINE
http://www.sciencemag.org/feature/plus/sfg/education/glossaries.shtml#postgenomics%20
Un eccellente elenco selezionato, classificato dai responsabili del sito, uno dei dieci "migliori" glossari online. Si vedano anche i glossari su tematiche correlate presso lo stesso sito.

GENOMICS GLOSSARY
http://www.genomicglossaries.com/
Una collezione di differenti glossari e tassonomie che include un glossario di bioinformatica.

HUMAN GENOME PROJECT INFORMATION GLOSSARY
http://www.ornl.gov/TechResources/Human_Genome/glossary/
Un utile glossario di termini genetici sviluppato dal DOE Human Genome Program, aperto alla navigazione e alla ricerca.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE (NHGRI) GLOSSARY OF GENETIC TERMS
http://www.genome.gov/glossary.cfm
Contiene contributi audio che permettono l'ascolto di definizioni e spiegazioni estese date da un esperto. In alcuni casi sono disponibili anche illustrazioni.

BASE DATI DI INTERESSE BIOINFORMATICO

I data base di sequenze e altri data base non-bibliografici sono le risorse di informazione più importanti nel settore della bioinformatica. In essi sono contenute un grande numero di informazioni correlate che si riferiscono ad un gene (posizione del genoma, struttura, sequenza, informazione sull'espressione chimica, ecc.) o a una proteina; questo implica una struttura della banca dati piuttosto complessa. I dati su sequenze che vengono versati in questi data base tendono a crescere esponenzialmente (si veda la crescita di Genbank

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)
Ci sono centinaia di base dati che possono essere considerati interessati per la bioinformatica: base dati specializzati per ogni specie, e basi dati distinti per i diversi tipi di informazione (sequenze di acido nucleico, sequenze di proteine, strutture di proteine, informazioni biochimiche e biofisiche ecc).
La seguente lista di banche ha lo scopo di elencare le base dati più importanti. Questa lista è molto selettiva e include solo pochi esempi rappresentativi di ogni tipo. L'enfasi è sui data base più grandi e più conosciuti, su quelli pubblici e gratuiti e su quelli che coprono dati sulla specie umana. Molti dei siti web forniscono a loro volta liste che vale la pena di consultare, come ad esempio:

NUCLEIC ACIDS RESEARCH: ANNUAL DATABASE ISSUE
http://www.nar.oupjournals.org/content/vol30/issue1/
Negli ultimi anni il journal Nucleic Acids Research, pubblicato dall'Oxford University Press, ha dedicato la prima uscita annuale ad elencare e descrivere i molti base dati di biologia molecolare e di bioinformatica. Questo numero include sempre molti articoli informativi che descrivono approfonditamente alcuni data base come anche una lista completa di base dati intitolato *Molecular Biology Database Collection* (http://www.nar.oupjournals.org/cgi/content/full/30/1/1/DC1).

DESCRIZIONI DI BASE DATI SRS (SEQUENCE RETRIEVAL SYSTEM)
http://downloads.lionbio.co.uk/publicsrs.html
Questo data base descrive l'elenco pubblico dei server SRS (http://downloads.lionbio.co.uk/publicsrs.html). Per esempio si veda il server dell' Istituto Europeo di Bioinformatica all'indirizzo:
http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page%2Btop%2B-newId.
Le banche dati sono elencate per tipo (ad es. librerie di sequenze, strutture 3D di proteine, mutazioni, SNP, traiettorie metaboliche, ecc).

BASE DATI DI SEQUENZE DI NUCLEOITIDICHE

GENBANK
http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html

ENTREZ NUCLEOTIDES DATABASE
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
GenBank è la base dati di sequenze nucleotidiche costruita e distribuita dal National Center for Biotechnology Information (NCBI) presso il National Institutes of Health. GenBank. Contiene più di 13 miliardi di basi da più di 100,000 specie e sta crescendo in modo esponenziale.

I dati vengono ottenuti tramite l'immissione diretta delle sequenze di dati da ricercatori, da progetti di sequenziamento in larga scala e dall'ufficio americano per i Brevetti.
Ci sono due modi di effettuare ricerche in GenBank: attraverso il sistema Entrez è possibile inoltrare query testuali tramite http://www.ncbi.nlm.nih.gov/Entrez/index.html, oppure si può inoltrare un query sequenza tramite la famiglia dei programmi BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). Il database è un insieme di sequenze provenienti da fonti diverse, tra cui GenBank, RefSeq, e la Protein Databank.
Le ricerche sul database Entrez Nucleotides si possono fare eseguendo query sui campi testuali e numerici del record, come il numero di accesso, la definizione, la parola chiave, il nome del gene e il nome dell'organismo ecc. Il numero di accesso è molto comodo poiché rappresenta un identificatore unico e coerente per la consultazione di GenBank nel suo insieme e non viene modificato neanche in presenza di successivi cambiamenti o aggiornamenti delle sequenze o del loro modo di essere annotate. I record delle sequenze nucleotidiche contenuti nel database nucleotidi sono collegati con la citazione PubMed dell'articolo in cui le sequenze sono state pubblicate. I record di sequenze di proteine sono collegate con la sequenza di nucleotidi da qui la proteina è stata tradotta. Sono inoltre presenti dei tutorial online dei database all'indirizzo: http://www.ncbi.nlm.nih.gov/Database/tut1.html
In genere, la maggior parte dei ricercatori, partendo da una sequenza di DNA di loro interesse, vogliono trovare le sequenze che più gli assomigliano. Ciò viene fatto tramite i programmi BLAST (Basic Local Alignment Search Tool). Il risultato è un report dettagliato che riporta la query, presenta una panoramica grafica delle similarità riscontrate nel database e descrive ogni allineamento significativo. Si possono apprendere altri dettagli sulle ricerche BLAST dalla pagina NCBI BLAST all'indirizzo:
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html.

BASE DATI DI SEQUENZE DI NUCLEOTIDI EMBL

http://www.ebi.ac.uk/embl/
La base dati di sequenze di nucleotidi EMBL costituisce la risorsa europea primaria per le sequenze di nucleotidi. Le principali fonti per sequenze di DNA e RNA sono sottomissioni dirette da ricercatori, progetti di sequenziamento del genoma e domande per brevetti. La base dati è prodotto attraverso una collaborazione internazionale con GenBank (USA) and the DNA Database of Japan (DDBJ). Ognuno di questi tre gruppi colleziona una porzione dei dati di sequenziamento totali riportati a livello mondiale e tutti i dati nuovi o aggiornati sono scambiati tra i gruppi giornalmente. Dalla home page del sito si possono sottomettere

ricerche di testo semplice alla base dati di sequenze di nucleotidi EMBL o alla Protein Databank (selezionando strutture proteiche dal menu) o alla base dati di sequenze proteiche Swall. Per ricerche più complesse, è consigliato accedere alla base dati attraverso il server SRS (Sequence Retrieval System). SRS è una modalità di eseguire query alla base dati o un sistema di navigazione dei dati simile al sistema Entrez. Permette la ricerca simultanea attraverso più base dati e di visualizzare i risultati della ricerca in modalità diverse. SRS può essere utilizzato per accedere a un grande numero di base dati includendo EMBL, SWISS-PROT e la Protein Databank, in base a la configurazione dello specifico server SRS che si utilizza.

Basi dati di Genomi

ENTREZ GENOME

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome%20

Il database Entrez Genome contiene l'intero genoma di oltre 800 organismi. I genoma rappresentano sia organismi interamente sequenziali che quelli per cui il sequenziamento è ancora in atto. Tutti e tre i principali domini della vita sono rappresentati – batteri, archaea, eukarioti – come anche molti virus e organelle. Ricerche testuali possono essere eseguite dalla pagina principale. I dati possono visualizzarsi anche alfabeticamente per specie: (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html), o gerarchicamente attraverso una lista tassonometrica, per arrivare a un insieme grafico del genoma di quell'organismo, poi a cromosomi specifici e ancora a geni specifici. A ogni livello ci sono: mappe e sommari, analisi appropriate per quel livello e collegamenti a record correlati da una varietà di altre base dati Entrez. È possibile fare ricerche con i programmi per la ricerca di similarità sulle sequenze genomiche.

Pagine molto utili per le specie più studiate (ad esempio gli umani, i topi, le mosche della frutta, i parassiti della malaria) possono essere trovate sulla pagina di Biologia Gnomica sotto "risorse per organismi specifici" (http://www.ncbi.nlm.nih.gov/Genomes/).

Queste pagine sono così dettagliate che ognuna potrebbe essere classificata come un sito web esaustivo. Ogni pagina raccoglie collegamenti ai dati gnomici, strumenti utili, fonti di dati correlati e informazioni sul genoma di quella specie specifica.

LA GUIDA AL GENOMA UMANO

http://www.ncbi.nlm.nih.gov/genome/guide/human/

è una buona sorgente di informazione.

HUMAN GENOME BROWSER DA UCSC
http://genome.ucsc.edu/
La sequenza del genoma umano è troppo grande per essere visualizzato tutto in una volta, quindi il Human Genome Browser è utile per una visualizzazione veloce di qualsiasi parte del menoma, richiesta a qualsiasi scala, con più di 24 tracce di informazioni (geni, EST, isole CpG, buchi di assemblaggio, banda cromosomica, …) associate alla sequenza completa del genoma umano.

THE GENOME DATABASE (GDB)
http://www.gdb.org/
The Genome Database è il contenitore centrale ufficiale per dati di mappatura gnomica provenienti dal Human Genome Iniziative. La base dati contiene tre tipi di dati: regioni del genoma umano, includendo geni, cloni e EST; mappe del genoma umano, includendo mappe citogenetiche, mappe di linkaggio, mappe ibride si radiazione, mappe di contenuti contig e mappe integrate (queste mappe possono essere visualizzate graficamente attraverso il web); variazione all'interno del genoma umano includendo mutazioni e polimorfismi, e dati di frequenze.
Ci sono opzioni per consultare geni contenuti in un cromosoma, geni mediante il nome di simbolo, mediante malattie genetiche. Ci sono molteplici modi per ricercare, comprese ricerche basate su testo per nomi di persone, per citazioni, per nomi di segmenti o per numeri di accessioni e ricerche di sequenze attraverso BLAST.

KEGG: KYOTO ENCYCLOPEDIA OF GENES AND GENOMES
Questa banca dati contiene informazioni sui pathway biochimici e un catalogo di geni. L'obiettivo primario di KEGG è di computerizzare la conoscenza attuale di interazioni molecolari quali pathway metabolici, pathway regolatori. Allo stesso tempo, KEGG mantiene un catalogo di geni per tutti gli organismi che sono stati sequenziali e collega ogni prodotto gene a un componente sul pathway. Poiché abbiamo bisogno di un catalogo in più di blocchi costruttivi, KEGG organizza una base dati di tutte le componenti chimiche delle cellule viventi e collega ogni componente al suo componente pathway.

BASE DATI DI SEQUENZE PROTEICHE

SWISS-PROT
http://www.expasy.org/
SWISS-PROT è una base dati di sequenze proteiche che ha lo scopo di fornire un alto livello di annotazioni (ad esempio la descrizione della funzione della proteina, la struttura del dominio della proteina, varianti, ecc.). I dati in Swiss-Prot derivano da traduzioni di sequenze DNA

dall'EMBL base dati di sequenze di nucleotidi, ricavati dalla letteratura o sottomessi direttamente dai ricercatori. Contiene annotazioni di alta qualità, è non-ridondante, ha referenze incrociate con altre base dati importanti quali l'EMBL base dati di sequenze nucleiche, PROSITE pattern database e PDB. Se si fa una ricerca tramite SWISS-PROT utilizzando SRS (Sequence Retrieval System) si possono eseguire ricerche più sofisticate e il formato dei risultati può essere scelto dall'utente. L'accesso a SWISS-PROT (sia direttamente sia tramite SRS) e a molti altri collegamenti ad altre risorse proteomiche sono fornite dal sistema ExPASy (Expert Protein Analysis System), il server di proteomica del Swiss Institute of Bioinformatics (SIB) http://www.expasy.org/

ENTREZ PROTEIN DATABASE

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?dB=Protein

La base dati di proteine contiene dati di sequenziamento da regioni di codifiche tradotte da sequenze DNA in GenBank, EMBL and DDBJ come anche le strutture proteiche sottomesse a PIR, SWISS-PROT, PRF, e il Protein Data Bank (PDB) (sequenze da strutture risolte). I record nativi di SWISS-PROT in genere contengono informazioni più dettagliate che quelle ottenute dai record dei base dati di proteine Entrez derivati dai record SWISS-PROT.

PROTEIN STRUCTURE DATABASE

PROTEIN DATA BANK (PDB)

http://www.rcsb.org/pdb/

Il PDB è stato attivato presso i Brookhaven National Laboratories nel 1971, cosa che ne ha fatto il primo database nella storia della bioinformatica. Il PDB è attualmente gestito dal Research Collaboratory for Structural Bioinformatics (RCSB) che è uno sforzo congiunto tra il San Diego Supercomputing Center, la Rutgers University e il National Institute of Standards and Technology (NIST). Il PDB è una raccolta di strutture tridimensionali di macromolecole biologiche (proteine, enzimi, acidi nucleici, complessi di acidi proteino-nucleici e virus) ottenute sperimentalmente, derivate da esperimenti in cristallografia a raggi x e NMR (per un'utile panoramica su questi metodi si veda http://www.rcsb.org/pdb/experimental_methods.html). L'immissione di strutture ottenute da modelli teoretici non è incoraggiata. I dati vengono immessi dalla comunità internazionale degli utenti e mantenuti dal personale dello RCSB PDB. Ogni settimana vengono inserite circa 50-100 nuove strutture. È disponibile una varietà di informazioni associate ad ogni struttura, tra cui "dettagli sulla sequenza, coordinate atomiche, condizioni di cristallizzazione, strutture 3-D di vicinato calcolate con diverse tecniche, dati geometrici derivati, fattori di strutture, immagini

3-D e una varietà di collegamenti ad altre risorse". Un metodo del sito NCBI Entrez è l'esecuzione di una ricerca di sequenza del tipo NCBI BLAST, con la selezione di "pdb" come database target. Diversi software possono essere impiegati per consultare i file PDB in 3-D, tra cui i browser plug-in RasMol e Chime e Deep-View.
Per ulteriori informazioni si vedano i tutorial sui PDB Query:
http://www.rcsb.org/pdb/query_tut.html
e la pagina di informazione e documentazione PDB:
http://www.rcsb.org/pdb/info.html#General_Information.

MMDB: MOLECULAR MODELING DATABASE

http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
MMDB è il database di strutture sviluppato presso l'NCBI.
Si tratta di un sottoinsieme di strutture tri-dimensionali ottenute dal Protein Databank (PDB). Lo MMDB si evidenzia per l'aggiunta di informazioni esplicite relative a grafi chimici e tramite il riferimento incrociato di dati strutturali a informazioni bibliografiche, ai database di sequenze e alla tassonomia NCBI. Il legame informativo esplicito produce una più coerente interpretazione dei dati delle coordinate tramite software di visualizzazione. MMDB può offrire dati per tre distinti visualizzatori di strutture: Cn3D, un viewer sviluppato dal NCBI; RasMol; MAGE.
Tutti sono disponibili per una varietà di sistemi operativi (Windows, MacOS, UNIX). La base dati di strutture può essere consultata direttamente usando numeri di accessione o elementi testuali quali nomi di autori, nomi di proteine, nomi di specie oppure date di pubblicazione. Il risultato condurrà a pagine di "query di strutture", che forniscono accesso a dati che corrispondono alle parole chiave.
Dalle pagine del sommario della struttura di un dato iniziale corrispondente si può accedere a sequenze di amino acidi e di acidi nucleici, recuperare documenti da PubMed, accedere a informazioni tassonometriche e lanciare il programma per visualizzare l'immagine 3D. La base dati strutture è notevolmente più piccola che le basi dati di proteine o nucleotidi di Entrez, ma un grande segmento di tutte le sequenze di proteine conosciute hanno omologhi in questo insieme di dati e si può spesso imparare di più su una proteina esaminando la struttura 3D del suo omologo. Sequenze di proteine sono collegate alle strutture 3D, quindi è possibile determinare se una struttura di proteine in Entrez ha omologhi tra le strutture conosciute esaminando le sue Sequenze Correlate o i 'Protein Neighbours' e controllando se questo insieme di dati ha dei collegamenti alle strutture relative.

Siti web utili alla Bioinformatica

ricerche in base dati e strumenti di analisi

http://www.ebi.ac.uk/Tools/index.html
Una lista di programmi che si possono usare tramite il web per sottomettere query alle basi dati di sequenze e per analizzarne i risultati. Questa lista proviene dall'European Bioinformatics Institute.

strumenti di proteomica expasy

http://www.expasy.org/tools/
Strumenti per la proteomica che possono essere utilizzati attraverso il web, coprono diverse categorie come ad esempio: identificazione e caratterizzazione di proteine, ricerche per similarità, previsione delle strutture secondarie, allineamento di sequenze.

Strumenti in linguaggi di programmazione

La bioinformatica fa uso di diverse lingue di programmazione quali C++, Perl, Java, Python, XML, Ruby and Lisp.
Vale la pena notare lo sviluppo di vari progetti Bio*.org che sono riuniti sotto il gruppo chiamato Open Bioinformatics Foundation (http://www.open-bio.org/), che è venuto in essere nell'ottobre del 2001. Ogni progetto è un'associazione internazionale di sviluppatori di strumenti open-source per la bioinformatica, la genomica e la ricerca nelle scienze della vita.

- BioPerl - http://bioperl.org/
- BioPython - http://www.biopython.org/
- BioJava - http://www.biojava.org/
- BioDAS - http://biodas.org/

Human Genome Project

Ci sono molti siti su questo argomento. Ecco tre siti di alto livello in termini di consultabilità e di collegamenti a ricerche di livello universitario:

- Genome Hub (dal National Human Genome Research Institute) http://www.genome.gov/
- The Human Genome (dal National Center for Biotechnology Information) - http://www.ncbi.nlm.nih.gov/genome/guide/human/
- Human Genome Project Information (dal U.S. Department of Energy) http://www.doegenomes.org/

Siti utili alle informazioni per lo studio dell'espressione dei Geni

Negli ultimi anni una nuova tecnologia, chiamata DNA microarray, ha sollevato grande interesse tra i biologi. Questa tecnologia promette il controllo dell'intero genoma come un singolo chip, cosicché i ricercatori possono avere in uno stesso momento una migliore idea delle interazioni tra migliaia di geni. Le tecniche impiegate includono, ma non si limitano a: biochip, DNA chip, DNA microarray e gene array.
La bioinformatica entra in scena negli esperimenti microarray in forma di elaborazione delle immagini, controllo sperimentale e analisi dei dati risultanti.
Sul sito presente presso EBI (http://industry.ebi.ac.uk/~alan/MicroArray/), sono presenti:

SMD microarray resources

http://genome-www4.stanford.edu/MicroArray/SMD/resources.html
Il sito presenta buoni contenuti ben organizzati. Riferimenti a database microarray, software, società e siti accademici. SMD Stanford Microarray Database, archivi dati e immagini microarray.

Database per la letteratura scientifica

PubMed - http://www4.ncbi.nlm.nih.gov/PubMed/
E' la più famosa collezione di letteratura biomedical. PubMed è l'interfaccia pubblica, il database per la letteratura medica (MEDLINE) prodotta dalla National Library of Medicine. Sono presenti oltre 11 milioni di citazioni MEDLINE di articoli e conferenze. PubMed è integrata con Entrez per fornire una completa possibilità di ricerca dei dati medici e genetici.

Organizzazioni importanti nell'ambito Bioinformatico

- American Crystallographic Association (ACA) http://www.hwi.buffalo.edu/ACA/
- Bioinformatics.org - http://bioinformatics.org/
- The Center for Information Biology and DNA Data Bank of Japan http://www.cib.nig.ac.jp/Welcome.html
- European Bioinformatics Institute (EBI) http://www.ebi.ac.uk/
- European Molecular Biology Laboratory (EMBL) http://www.embl.de/
- Federation of American Societies for Experimental Biology (FASEB) http://www.faseb.org/
- The Human Genome Organisation (HUGO) http://www.gene.ucl.ac.uk/hugo/
- International Society for Computational Biology (ISCB) http://iscb.org/
- National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/

- National Center for Genome Resources http://www.ncgr.org/
- National Human Genome Research Institute (NHGRI) http://www.nhgri.nih.gov/
- National Library of Medicine (NLM) http://www.nlm.nih.gov/
- Swiss Institute of Bioinformatics (SIB) http://www.isb-sib.ch/

Sintesi

E-SCIENCE BIOINFORMATICS

The success of human genome sequencing and the hope for a new era, when genomic research may improve human condition, boosts huge interest in bioinformatics, a new discipline in between molecular biology and informatics. In this area, international initiatives have been activated for information gathering, as well as programmes and scientific literature useful for researches in the field of Bioinformatics.

In this paper, the most important terms, concepts and resources of bioinformatics will be defined. Bioinformatics may be defined as follows: "Bioinformatics is the scientific field where biology and information & communication technology (ICT) meet".

In this field, Bioinformatics deals with acquisition, memorization, distribution, analysis and interpretation of data, mainly in the field of molecular biology, genetics and biochemistry, with more and more important links with medicine.

Bioinformatics aims at providing investigation methodologies and information quickly, so that they permit, for example, to support medical science with information necessary to understand the basic mechanisms of all possible disorders. Therefore, methods and tools made available by Bioinformatics play a fundamental role in the development of biotechnologies.