UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scuola di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica

APPLICAZIONI DELL'ALGORITMICA ALLA BIOLOGIA: EVOLUTIONARY TREES E CLUSTERING

APPLICATIONS OF ALGORITHMICS TO BIOLOGY: EVOLUTIONARY TREES AND CLUSTERING

MATTEO TORTOLI

MARIA CECILIA VERRI

Anno Accademico 2018-2019

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1

CAPITOLO 1: CONCETTI BASE DI BIOLOGIA

La bioinformatica è una materia che tratta tanto l'informatica quanto la biologia, pertanto è necessario illustrarne gli argomenti più importanti, che verranno spiegati in modo funzionale allo scopo della presente tesi.
La *biologia* è la scienza che studia la vita, dagli attori che ne fanno parte fino ai processi in cui essi sono coinvolti. Poiché la vita nella terra si estende dalla biosfera fino alle molecole, si è resa necessaria l'esigenza di trovare una vera e propria scala a livello globale. Tra le varie forme di vita si trovano, omettendo quelle non interessate ed in ordine crescente, le molecole (insiemi di atomi), le macromolecole (insieme di molecole) e le cellule (insieme di macromolecole).
Ci sono quattro tipi di macromolecole che risultano essenziali per tutte le forme di vita:

- *Polisaccaridi*: macromolecole formate da un'insieme molecole, ovvero i monosaccaridi, tra cui il fruttosio, il glucosio e così via.
- *Proteine*: sono la "struttura" degli esseri viventi, infatti consentono lo sviluppo e mantenimento degli organi.
- *Lipidi*: chiamati anche grassi, sono le riserve di energia.
- *Acidi nucleici*: DNA e RNA.

Di seguito vengono approfonditi il DNA e l'RNA.

1.1 DNA

Il *DNA* o *acido desossiribonucleico* è una macromolecola contenente il patrimonio genetico degli esseri viventi. Il patrimonio genetico contiene tutte le informazioni genetiche[1] di un organismo.
Di seguito viene illustrata un'immagine del DNA, insieme ad una breve introduzione.

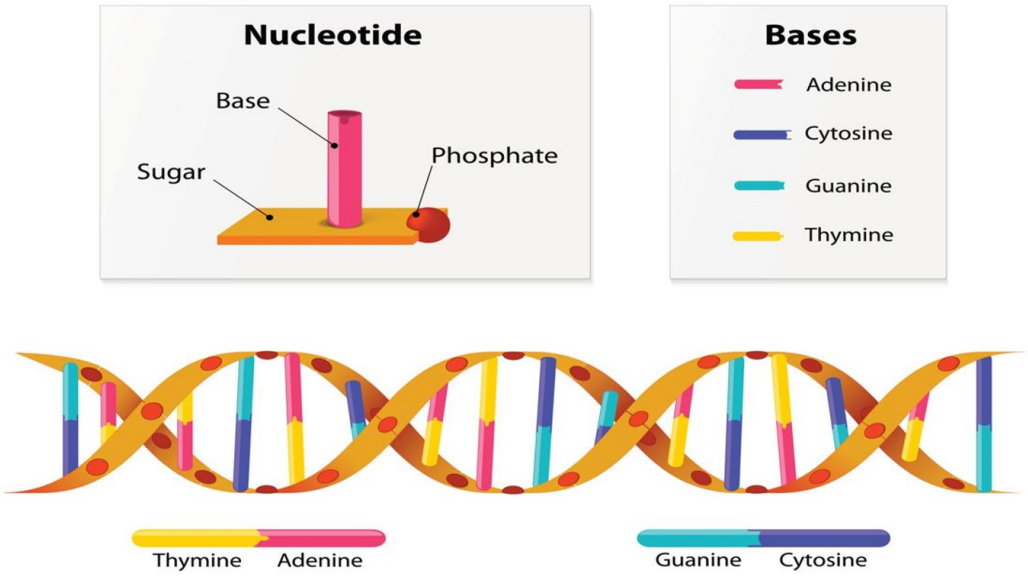

Figure 1: La struttura del DNA e del nucleotide.

La struttura del DNA è ottimizzata per contenere informazioni, caratterizzata da una doppia elica dove ciascun filamento è formato da una *sequenza di nucleotidi* dalla lunghezza variabile, che possono essere considerati i suoi "mattoncini".
Ciascun nucleotide è composto da una molecola di zucchero, un gruppo fosfato[2] ed una base azotata. Queste ultime legano assieme i due filamenti del DNA, e è possibile trovarne di quattro tipi:

- *Adenina* (A)
- *Timina* (T)
- *Guanina* (G)

1 successivamente verrà mostrata la definizione di gene.
2 Il gruppo fosfato è un un insieme di elementi strutturati e ben definiti

- *Citosina* (C)

La base azotata Adenina si può legare solo la Timina (A-T), mentre la Guanina solo con la Citosina (G-C), questo significa che i filamenti sono complementari, e quindi se conosciamo le basi di un filamento sappiamo pure quelle dell'altro filamento(se una base azotata di un filamento è la Timina, allora la corrispondente base azotata dell'altro filamento è l'adenina). TODO: descrivere brevemente il gene.

1.2 RNA

L'*RNA*, è una macromolecola che sta per acido ribonucleico, caratterizzato da una struttura a singolo filamento composto da una *sequenza di ribonucleotidi* più o meno lunga.
I ribonucleotidi si differenziano dai nucleotidi per una differente molecola di zucchero e perché la Timina (T) è sostituita con l'Uracile(U).

2

CAPITOLO 2: LA BIOINFORMATICA

Per molti anni l'informatica è stata una scienza a sé stante, tuttavia negli ultimi decenni, grazie al progresso scientifico e tecnologico, sono nate nuove discipline chiamate genericamente **X-Informatics**. Queste sono il risultato dell'incontro tra l'informatica ed altre scienze di base (quali la biologia, la chimica, l'astronomia, la geologia etc) e tra queste citiamo la bioinformatica, la chemioinformatica, l'astroinformatica, la geoinformatica e così via. Anche se queste discipline sono diverse tra loro, queste condividono gli stessi obiettivi:

- Processamento ed estrazione delle informazioni
- Interazione con la raccolta dati adattata e personalizzata per l'utente
- Integrazione dei dati ottenuti tra sorgenti eterogenee
- Utilizzo trasparente ed efficiente dei dati in base al contesto scientifico, dalla raccolta, all'analisi, fino alla catagolazione
- Fornire supporto decisionale per l'utente, riducendo così i possibili errori e facilitando l'analisi dei risultati

Tra tutte queste discipline risulta di particolare importanza la bioinformatica.

2.1 CHE COSA È LA BIOINFORMATICA?

Non esiste un unico modo con cui definire la bioinformatica, infatti è possibile trovare definizioni diverse tra loro in quanto i professionisti non sempre concordano sulla portata del suo uso, sia nel campo della biologia che dell'informatica. Tuttavia una possibile definizione è la seguente:

La bioinformatica è un campo multidisciplinare della scienza che coinvolge la genetica, la biologia molecolare, l'informatica, la matematica e la statistica,

rivolta a studiare sistemi biologici utilizzando metodi e modelli informatici e computazionali.

Tra i vari obiettivi precedentemente elencati, va aggiunto quello che risulta essere l'obiettivo principale di questa disciplina, ovvero aumentare la conoscenza di tutti quei processi ed attori presenti in natura.
A prescindere dal problema in questione, è possibile individuare un approccio standard, suddiviso in quattro step:

- Analisi del problema da affrontare
- Collezionamento ed analisi di dati statistici ottenuti a fronte di dati biologici in input
- Creazione di modelli ed uso di strumenti matematici che possano essere applicati al problema in esame, al fine di sviluppare un algoritmo
- Creazione, valutazione e test dell'algoritmo risolutivo del problema

Una parte fondamentale della bioinformatica consiste in esperimenti che generano dati ad alto throughput (high-throughput data), tra cui la misurazione dei modelli di espressione genica oppure la determinazione della sequenza nucleotidica nel caso del DNA e RNA e aminoacidica nel caso delle proteine. Per *high-throughput data* si intendono tutti quei dati biologici ottenuti tramite tecniche automatizzate e quindi non ottenibili attraverso metodi convenzionali. Il mining di questi dati può portare a nuove scoperte scientifiche non solo in campo biologico, ma anche medico, sia nel breve che nel lungo periodo. Ad esempio grazie all'estrazione dei dati ottenuti dal *progetto genoma umano*[1], sono stati scoperti nuovi geni legati alle malattie e nuovi bersagli molecolari[2]. Nel lungo periodo sarà possibile scoprire eventuali reazioni avverse ai farmaci che sono differenti da individuo ad individuo, al punto tale che, grazie all'informazione genetica ottenuta attraverso strumenti informatici, sarà possibile personalizzare l'uso di farmaci, portando ad una migliore efficacia della terapia individuale, riducendo o addirittura eliminando possibili effetti collaterali.

1 Il progetto genoma umano (Human Genome Project) è stato uno dei più grandi progetti scientifici degli ultimi anni. L'obiettivo era quello di ottenere la sequenza del genoma umano (e quindi il suo intero DNA) e identificare i geni contenuti in esso. Il progetto è cominciato nel 1990, per poi essere completato nel 2003 ed ulteriori ricerche sono ancora in corso.

2 Il bersaglio molecolare è una qualsiasi proteina o enzima su cui si può intervenire per modificare il decorso di una malattia.

2.2 STORIA

Fino agli anni '50 il DNA era ancora una scoperta a "metà", il World Wide Web non era ancora nato e nessuno sentiva l'esigenza di dover creare algoritmi per analizzare e memorizzare i dati biologici. Tuttavia le cose cambiarono dal 1953 in poi, con la scoperta della struttura a doppia elica del DNA. Da quel periodo in poi si sono susseguite una serie di scoperte scientifiche che hanno reso la biologia una scienza molto orientata ai dati. La bioinformatica nasce verso la fine degli anni '70, con la scoperta delle prime sequenze nucleotidiche del DNA, poiché comincia a nascere l'esigenza di poter archiviare i dati e consultarli quando necessario. Da quegli anni in poi la bioinformatica è cresciuta insieme alla biologia e tuttora è una scienza in continua evoluzione.
È possibile trovare altre due date importanti nella storia della bioinformatica:

- *Anno 1990*: Data di creazione del *Basic Local Alignment Search Tool (BLAST)*[5] ovvero un algoritmo che permette il confronto tra sequenze nucleotidiche e aminoacidiche
- *Anno 2003*: Completamento del *progetto Genoma Umano*, che ha permesso la scoperta dell'intero patrimonio genetico dell'essere umano

2.3 AREE DI RICERCA

Data la natura eterogenea dei dati biologici, la bioinformatica comprende un vasto numero di aree di ricerca in continua crescita. Di seguito verranno elencate le principali, assieme ai relativi algoritmi più importanti. Sarà possibile notare che alcuni degli algoritmi presentati vengono usati in aree di ricerca diverse, grazie alla loro scalabilità.

2.3.1 *Analisi dei genomi*

Uno dei principali focus della bioinformatica riguarda l'analisi dei genomi[3] degli organismi il cui sequenziamento[4] è già stato completato, dal moscerino

3 Per genoma si intende l'intero materiale genetico di un organismo, composto da DNA o RNA.

4 Il sequenziamento è un processo mediante il quale viene determinata la struttura primaria delle macromolecole del DNA, RNA (sequenze di nucleotidi) e proteine(sequenze di aminoacidi).

della frutta fino all'essere umano. L'analisi dei genomi è un'area di ricerca che riguarda non solo la bioinformatica, ma anche la genomica, ovvero quella disciplina che studia la struttura, il contenuto, la funzione e l'evoluzione del genoma.
Perché analizzare i genomi? Un gene viene sequenziato per conoscere la sua funzione ed eventualmente per poterla modificare. La conoscenza dell'intero genoma di un organismo fornisce le sequenze di tutti i suoi geni, permettendo così di identificare e manipolare quelli che influenzano il metabolismo, lo sviluppo cellulare e i processi patologici negli esseri umani, animali e nelle piante.
L'obiettivo dell'analisi dei genomi è di identificare e modificare i geni che abbiano una particolare funzione biologica attraverso strumenti computazionali, ovvero quelli che permettono la risoluzione di problemi altrimenti inaccessibili con i tempi e le modalità umane.

2.3.2 *Analisi di sequenze*

L'analisi delle sequenze di DNA, RNA o proteine è un processo mediante il quale tali macromolecole vengono sottoposte a dei metodi analitici al fine di capirne la struttura e le funzionalità.
Di particolare importanza risulta lo studio delle sequenze di DNA[5] che possono essere memorizzate in un computer attraverso una vasta varietà di metodi. La memorizzazione avviene attraverso l'uso di caratteristiche identificative di una determinata sequenza di DNA, ad esempio dando un nome al gene o indicandone la fonte, dopodiché vengono salvate all'interno di database che prendono il nome di *database biologici* (vedere sottosezione 2.3.10).
Grazie all'analisi delle sequenze genetiche, è possibile individuare le mutazioni di geni alla base di potenziali malattie.
Una volta estratto il DNA, vengono create migliaia e migliaia di copie di un singolo frammento e successivamente inserite in macchinari chiamati *DNA Sequencer*. Queste ultime svolgono il sequenziamento del DNA in modo automatico. infine i dati vengono raccolti ed analizzati.
Tra i vari algoritmi utilizzati per l'analisi delle sequenze risultano di particolare importanza gli *algoritmi di Clustering*, il cui obiettivo è quello di raggruppare i dati delle sequenze in modo veloce e preciso. I cluster giocano un ruolo chiave all'interno della bioinformatica, infatti data la

5 Il sequenziamento del DNA consiste nel determinare la sequenza di nucleotidi all'interno di un suo frammento.

grande quantità di dati da gestire, è possibile raggrupparli in insiemi (cluster, appunto) secondo determinati criteri, di modo che gli elementi che sono simili tra di loro siano nello stesso cluster mentre quelli che sono differenti risiedono in altri. I principali algoritmi di Clustering nella bioinformatica sono il *Clustering gerarchico* e l'*algoritmo k-Means*, che verranno spiegati successivamente.

2.3.3 *Analisi dell'espressione genica*

Quando un gene è attivo, si intende dire che è "espresso", ovvero che è stato prima trascritto in una copia di mRNA (RNA messaggero) e poi tradotto in proteina, con questo concetto si intende *espressione genica*. Questo vuol dire che esistono geni che non vengono mai espressi e pertanto
L'analisi dell'espressione genica quantifica l'attività dell'espressione di migliaia di geni simultaneamente, per capire in quali condizioni alcuni di questi risultano attivi ed altri no.
In questa area di ricerca gli algoritmi di data mining e di Clustering giocano un ruolo di fondamentale importanza, in quanto i primi consentono di estrarre le informazioni mentre i secondi permettono di classificarle in gruppi (in particolar modo il Clustering gerarchico e l'algoritmo K-means).
Ma perché l'analisi dell'espressione genica è importante? Le motivazioni sono molteplici. Prima di tutto, se l'espressione di un gene non noto è simile a quella di un gene conosciuto, è possibile che essi abbiano funzioni simili o che siano coinvolti nello stesso meccanismo biologico. Questo può portare ad una sorta di "reazione a catena", dove partendo da un gene conosciuto, si scoprono altre funzioni simili di geni non conosciuti. In secondo luogo, è importante anche nel settore della biomedicina, predicendo eventuali metastasi tumorali.

2.3.4 *Analisi ed elaborazione di bioimmagini*

L'analisi ed elaborazione delle bioimmagini consiste nell'usare metodi informatici per acquisire, analizzare, fare data mining di immagini ottenute microscopio, con l'obiettivo di risolvere problemi di natura biologica e medica. Questa area di ricerca si basa principalmente sul machine learning, in particolar modo sul *pattern recognition*, ovvero lo studio di come le macchine possano imparare a distinguere vari modelli e a prendere delle

decisioni in base a specifici pattern, ponendosi come obiettivo principale, nel caso della bioinformatica, la classificazione di organismi.

2.3.5 *Filogenetica*

La filogenetica è quel campo della biologia evolutiva[6] che studia le relazioni evolutive tra vari gruppi di organismi. Tradizionalmente si basava sul confronto tra le caratteristiche fenotipiche degli organismi, oggi invece si usano i dati ottenuti tramite sequenziamento genico, permettendo la costruzione di alberi filogenetici estremamente accurati.
L'albero filogenetico è un diagramma che rappresenta le relazioni evolutive tra i vari organismi, la cui invenzione è dovuta a Charles Darwin, nel 1837. Una delle peculiarità di questi diagrammi è la possibilità di costruirli in base a dati genetici, genomici o morfologici, al fine di descrivere le relazioni che vi sono tra organismi viventi oppure tra specie estinte e specie viventi. Con questa nuova informazione è possibile dare una definizione alternativa alla filogenetica, ovvero come quella disciplina che si occupa di costruire ed analizzare gli alberi filogenetici.
Le applicazioni della filogenetica sono molteplici, elencate di seguito.

- *Bioinformatica e computing*: Gli alberi risultano una struttura dati di fondamentale importanza nel campo dell'informatica e degli algoritmi, infatti, molti di questi sviluppati per la filogenetica sono stati successivamente utilizzati anche in altri settori
- *Classificazione*: fornisce dei metodi di classificazione degli organismi viventi in modo accurato
- *Medicina forense*: può essere usata per valutare delle prove di DNA in casi giudiziari
- *Identificazione dell'origine evolutiva degli agenti patogeni*: è possibile studiare ancora più a fondo gli agenti patogeni[7], prevenendo eventuali epidemie. Infatti, scoprire a quale specie vivente è collegato un determinato agente patogeno fornisce delle informazioni per scoprire quale potrebbe essere una eventuale forma di trasmissione
- *Conservazione delle specie*: contribuisce ad impedire l'estinzione di specie animali e vegetali

6 La biologia evolutiva si occupa dello studio delle origini ed evoluzione delle specie.
7 Gli agenti patogeni sono i virus,i batteri, etc...

2.3.6 *Bioinformatica strutturale*

Le macromolecole, tra cui DNA, RNA e proteine, svolgono la loro funzione all'interno dei sistemi biologici grazie alla loro conformazione tridimensionale, cioé una forma particolare che assumono nello spazio: questo è particolarmente vero per le proteine che senza specifici ripiegamenti su sé stesse sono prive di qualsiasi funzione. Tuttavia conoscere tale struttura non è affatto facile, basti pensare che in natura esistono 20 diversi tipi di aminoacidi e che se prendiamo una proteina composta da una sequenza di 70 di questi, otteniamo $20^{70} (= 1.180591620717411303424 \times 10^{91})$ possibili strutture (anche se la natura non ne ha selezionate così tante!). Ed è qui che entra in gioco la bioinformatica strutturale, ovvero quella area di ricerca che si pone l'obiettivo di analizzare e ricostruire, tramite algoritmi, la struttura tridimensionale delle macromolecole.
Grazie a questa disciplina è possibile conoscere le interazioni fra macromolecole, fornire dati biologici aggiuntivi e predire[8] la loro struttura tridimensionale, permettendo quindi di conoscerne le funzionalità, in particolar modo per le proteine. Proprio quest'ultimo obiettivo risulta di particolare importanza non solo per la bioinformatica ma in generale per la biologia stessa, oltre ad essere tutt'ora una grande sfida per la scienza.
In passato sono stati utilizzati vari approcci di tipo computazionale, tra cui gli algoritmi evolutivi, tuttavia non risultavano particolarmente efficaci per questo tipo di problema, al contrario invece degli algoritmi genetici.
Gli *algoritmi genetici* sono metodi complessi che hanno il compito di risolvere problemi che si basano sulla selezione naturale, in altre parole, dato un punto di partenza, cercano di scegliere le soluzioni migliori e ricombinarle tra di loro per ottenere una soluzione "ottima". Data una popolazione in input, ad ogni iterazione vengono scelti casualmente degli individui, chiamati genitori, che verranno usati per creare altri individui, chiamati figli, che a loro volta verranno utilizzati nella iterazione successiva, di modo che con il passare delle generazioni si raggiunge la soluzione ottima.
Questi algoritmi risultano particolarmente efficienti nella bioinformatica strutturale per due motivi, il primo è perché riescono a risolvere problemi

8 Con predizione si intende predire ogni atomo di cui è composta la macromolecola nelle tre dimensioni.

complessi velocemente, sfruttando la parallelizzazione automatica[9] ed il secondo perché sono particolarmente ottimizzati per la ricerca genetica. Da citare *Proteine Data Bank*[23], un vero e proprio archivio di acidi nucleici e proteine visualizzabili in 3-D, che risulta di fondamentale importanza in questa area di ricerca.

2.3.7 *Genetica delle popolazioni*

Prima di poter definire che cosa è la genetica delle popolazioni, è necessario introdurre alcuni concetti, di seguito elencati.

- *Popolazione*: è un gruppo di organismi (vegetali e non) che vivono nello stesso luogo e che condividono determinate proprietà biologiche, pertanto sono della stessa specie
- *Alleli*: ogni gene esiste in due o più possibili "versioni", grazie ai suoi alleli. Gli alleli sono le diverse sequenze possibili per un gene, pertanto la loro combinazione determina il suo carattere ereditario (colore degli capelli, degli occhi, ecc...). Ad esempio, nel caso del gene del colore di un fiore, ci può essere un allele per il colore rosso ed un altro per il giallo. Se il primo risulta dominante, allora il fiore sarà di colore rosso
- *Frequenza genica*: misura la frequenza di un allele di un determinato gene presente in una popolazione
  Preso in considerazione un gene di una popolazione, infatti, è possibile trovare alleli diversi con frequenze diverse.

La genetica delle popolazioni, quindi, si occupa di studiare la frequenza dei geni delle popolazioni e della loro variazione nello spazio e nel tempo, dalla loro origine fino alla loro evoluzione.
Poiché in questa area di ricerca vengono coinvolti i processi evolutivi degli organismi viventi, proprio come nella bioinformatica strutturale, gli algoritmi genetici giocano un ruolo chiave (già definiti nella sottosezione 2.3.6).
È possibile dare una sguardo più approfondito a tali algoritmi, individuando cinque fasi fondamentali:

9 La parallelizzazione è un processo mediante il quale invece di eseguire un task alla volta, viene spezzettato in più "sotto-task" indipendenti, che quindi possono essere eseguiti in contemporanea.

1. *Popolazione iniziale*: data una popolazione con un determinato problema, vengono scelti un gruppo di geni, rappresentati tramiti stringhe dell'alfabeto
2. *Funzione di fitness*: funzione che associa ad ogni individuo di una popolazione un punteggio che varia in base alla sua abilità nel competere con altri individui
3. *Selezione*: vengono selezionati gli individui con il punteggio della funzione di fitness migliore, affinché passino i propri geni alla generazione successiva
4. *Incrocio*: per ogni coppia di genitori viene scelto una sorta di punto di scambio tra i loro geni, di modo che i geni dei figli saranno il risultato dell'incrocio tra i geni dei suoi genitori. Questa risulta essere la parte più importante di tali algoritmi
5. *Mutazione*: in alcuni casi ci possono essere delle mutazioni all'interno dei geni dei figli
6. *Terminazione*: l'algoritmo termina quando non verranno generati più figli, in quanto il problema di partenza è stato risolto

2.3.8 *Biologia dei sistemi*

Un sistema biologico è una vera e propria rete di entità biologiche connesse tra di loro, ad esempio, il sistema nervoso di un essere umano è un sistema biologico, composto da un'insieme di entità, ovvero il midollo spinale, i nervi, il cervello ed il cervelletto.
La biologia dei sistemi, quindi, è quella area di ricerca che si occupa di studiare i sistemi biologici attraverso metodi computazionali e modelli matematici e statistici. L'approccio alla materia di studio può essere bottom-up, e quindi si inizia dallo studio dei geni come sistemi biologici, proteine, etc, unendo di volta in volta queste entità, arrivando all'organismo nella sua interezza, ma anche al contrario, e quindi un approccio top-down.
Le applicazioni della bioinformatica in questa area di ricerca sono molteplici: l'ottenimento di dati ad alto throughput (high-throughput data), ottenuti attraverso tecniche di analisi statistica avanzate, ma soprattutto la creazione di un linguaggio di Markup per i sistemi biologici, ovvero il *Systems Biology Markup Language* (SBML)[27]. Tale linguaggio è nato con lo scopo di rappresentare e modellare i sistemi biologici (e non solo)

attraverso l'uso delle macchine, grazie anche ad un buon numero di tools nati per supportare ed espandere il suo uso, tra cui il framework *Systems Biology Workbench* (SBW)[25]. Il SBW è un insieme di strumenti che permettono di creare, visualizzare e simulare le reti tra le entità che compongono un sistema biologico.

2.3.9 *Data mining*

Data mining, chiamata anche *Knowledge Discovery* è una tecnica di estrazione di dati non conosciuti e potenzialmente utili a fronte di una grande mole di dati. Con il passare del tempo e con il crescere della quantità di dati biologici, data mining sta assumendo sempre di più un ruolo centrale, questo si traduce in molte applicazioni, tra cui la ricerca di dati per predire e conoscere la struttura tridimensionale delle proteine, la scoperta delle "cause genetiche" alla base di una malattia e la facilitazione dell'uso degli algoritmi di clustering di sequenze.
I temi principali del data mining nella bioinformatica sono tre, di seguito elencati.

1. *Data Preprocessing e Data Cleaning*: data l'eterogeneità dei dati biologici, è necessario eliminare dal set di dati da controllare tutti quelli che risultano formalmente errati (ad esempio "tipo Macromolecola: RNA, Base: Timina") e quelli che risultano inconsistenti, corrotti o mancanti

2. *Data mining tools*: con il passare degli anni sono stati creati molti strumenti per l'analisi dei dati biologici, quindi, una volta preprocessati e puliti, si possono analizzare grazie proprio a questi strumenti, quali ad esempio *GeneSpring*

3. *Nuovi metodi di data mining*: nuove ricerche scientifiche richiedono nuovi metodi di estrazione dei dati. Tali metodi devono essere efficienti e scalabili, al fine di poter analizzare al meglio i dati biologici

2.3.10 *Database biologici*

Poiché la biologia, con il passare degli anni, è diventata una scienza ricca di dati, è nata l'esigenza di catalogarli e memorizzarli. È proprio da ciò che sono nati i database biologici, ovvero quella collezione di

informazioni che comprendono sia pubblicazioni scientifiche che dati provenienti da ricerche, di seguito elencati.

- Sequenze nucleotidiche (DNA e RNA)
- Sequenze di aminoacidi (proteine)
- Informazioni su geni (dalle espressioni geniche al progetto Genoma Umano)
- Modelli in tre dimensioni delle proteine (ad esempio il Proteine Data Bank[23])

Poiché questi dati risultano non solo molto diversi tra loro, ma anche complessi da gestire sia nella loro individualità che in relazione agli altri dati (data la loro natura articolata), la loro modellazione risulta un punto cardine per la creazione di database biologici, nel quale è possibile individuare tre concetti fondamentali.

1. *Ordine delle sequenze*: Sia le sequenze nucleotidiche del DNA e RNA che le sequenze di aminoacidi delle proteine possono essere modellate come entità statiche. Questo è dovuto principalmente al fatto che sono delle proprietà interne ad altre entità biologiche e che le sequenze cambiano con il passare del tempo, ma solo leggermente (in base all'evoluzione).
   Questo concetto è molto importante nella modellazione delle sequenze, in quanto un piccolo cambiamento può avere un grosso impatto sulle entità più grandi

2. *Processi di input/output*: Poiché nella biologia ci sono molti processi che, partendo da un input, restituiscono un output, è necessario identificare ed etichettare il ruolo di queste entità. Ad esempio basti pensare che, partendo con delle sequenze nucleotidiche del DNA, si ottengono delle sequenze di aminoacidi di proteine, grazie ai processi di trascrizione e traduzione.
   Anche questo concetto risulta avere un ruolo chiave per poter trattare i dati biologici in modo dinamico

3. *Relazioni spaziali tra le molecole*: come già detto nella sottosezione "bioinformatica strutturale", la struttura in tre dimensioni delle macromolecole (in particolar modo le proteine) risulta di fondamentale importanza per comprenderne le funzioni, pertanto, a livello di database, è necessario definire le relazioni tra le varie entità

che compongono una macromolecola, di modo che grazie a queste informazioni sia possibile ricostruire il suo modello in 3-D.

I database biologici vengono suddivisi in due categorie, ovvero *di primo livello* e *specializzati*. I primi contengono solamente le entità statiche, e quindi le sequenze nucleotidiche e di aminoacidi, mentre i secondi raccolgono informazioni relative alle loro funzioni, alle malattie dovute a mutazioni, pubblicazioni scientifiche, e così via.
Tra i database biologici più famosi è possibile trovare Ensembl[13], National Center for Biotechnology Information (NCBI)[22] ed il già citato Proteine Data Bank[23].

BIBLIOGRAPHY

[1] Allegretti M. (2014).
La biologia strutturale in movimento.
AIRInforma: AIRIcerca.
http://informa.airicerca.org/it/2014/10/04/biologia-strutturale-in-movimento/

[2] Bacic, A., Likić, V. A., Lithgow, T., McConville, M. J.(2010).
Systems Biology: The Next Frontier for Bioinformatics.
Advances in Bioinformatics, 2010, 1–10.
DOI:
l10.1155/2010/268925

[3] Bateman, A., Peng, H., Valencia, A., Wren, J. D. (2012).
Bioimage informatics: a new category in Bioinformatics.
Bioinformatics, 28(8), 1057–1057.
DOI:
10.1093/bioinformatics/bts111

[4] Bayat Ardeshir (2002).
Science, medicine, and the future: Bioinformatics.
BMJ, 324(7344), 1018–1022. DOI:
10.1136/bmj.324.7344.1018

[5] Basic Local Alignment Search Tool Website.
BLAST.
https://blast.ncbi.nlm.nih.gov/Blast.cgi (Cited on page 13.)

[6] Borne, K. D.
X-Informatics: Practical Semantic Science.
http://adsabs.harvard.edu/abs/2009AGUFMIN43E..01B.
George Mason University Fairfax VA, American Geophysical Union Fall Meeting, 12/2009.

[7] Breitling, R. (2010).
What is systems biology? Frontiers in Physiology.
DOI:
10.3389/fphys.2010.00009

[8] Candavelou, M., Gollapalli, S., Gopal,J., Karthikeyan, K., Venkatesan, A.(2013).
Computational Approach for Protein Structure Prediction.
Healthcare Informatics Research, 19(2), 137.
DOI:
10.4258/hir.2013.19.2.137

[9] Can T. (2013).
Introduction to Bioinformatics.
miRNomics: MicroRNA Biology and Computational Analysis.
DOI:
10.1007/978-1-62703-748-8_4

[10] Charette, S. J., Derome, N., Gauthier, J., Vincent, A. T.(2018).
A brief history of bioinformatics.
Briefings in Bioinformatics.
DOI:
10.1093/bib/bby063

[11] Chen J., Sidhu S. A.
Biological Database Modeling.
Norwood, MA, Artech House, 2008.

[12] Emery L.
Phylogenetics: An introduction.
European Bioinformatics Institute-European Molecular Biology Laboratory.
https://www.ebi.ac.uk/training/online/course/introduction-phylogenetics
Oxford.

[13] Ensembl Website.
Ensembl.
http://www.ensembl.org/ (Cited on page 22.)

[14] Haque Sultan Omar (2016).
Phylogenetics.
Enciclopedia Britannica.
https://www.britannica.com/science/phylogenetics#accordion-article-history

[15] ICAR CNR: Istituto Di Calcolo E Reti Ad Alte Prestazioni.
Bioinformatica.
https://www.icar.cnr.it/bio-informatica/

[16] Jiang, X., Lin, G., Liu, X., Zeng, X., Zou, Q.(2018).
Sequence clustering in bioinformatics: an empirical study.
Briefings in Bioinformatics.
DOI:
https://doi.org/10.1093/bib/bby090

[17] Jones C. Neil and Pevzner A. Pavel.
An introduction to bioinformatics algorithms.
Massachusetts, Massachusetts Institute of Technology, 2004.

[18] MathWorks.
What Is the Genetic Algorithm?
https://it.mathworks.com/help/gads/what-is-the-genetic-algorithm.html

[19] Mallawaarachchi V (2017).
Introduction To Genetic Algorithms-Including Example Code.
https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3

[20] Mount W. David.
Bioinformatics: Sequence and Genome Analysis.
Cold Spring Harbor Laboratory Press, 2004.

[21] National Human Genome Research Institute (2015).
DNA Sequencing Fact Sheet.
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet

[22] National Center for Biotechnology Information Website.
NCBI.
https://www.ncbi.nlm.nih.gov/ (Cited on page 22.)

[23] Proteine Data Bank Website.
PDB.
http://www.rcsb.org/ (Cited on pages 18, 21, and 22.)

[24] Raut, A., Raut, S. A., Sathe, S. R.(2010).
Bioinformatics: Trends in gene expression analysis.

International Conference on Bioinformatics and Biomedical Technology.
DOI:
10.1109/icbbt.2010.5479003

[25] Systems Biology Workbench Website.
SBW.
http://sbw.sourceforge.net/ (Cited on page 20.)

[26] Semple C., Steel M.
Phylogenetics.
Oxford, Oxford University Press, 2003.

[27] Systems Biology Markup Language.
SBML.
http://sbml.org/ (Cited on page 19.)

[28] Templeton R. Alan.
Population Genetics and Microevolutionary Theory.
Washington University, St. Louis, Missouri, Wiley-Liss, 2006.

[29] Tramontano Anna (2003).
La grande scienza. Bioinformatica.
http://www.treccani.it/enciclopedia/la-grande-scienza-bioinformatica_%28Storia-della-Scienza%29/

[30] Dennis E. Shasha, Hannu Toivonen, Jason T. L. Wang, Mohammed J. Zaki.
Data Mining in Bioinformatics.
London, Springer, 2004.