UNIVERSITÀ
DEGLI STUDI
FIRENZE

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

APPLICAZIONI DELL'ALGORITMICA ALLA BIOLOGIA: ALBERI EVOLUTIVI

APPLICATIONS OF ALGORITHMICS TO BIOLOGY: EVOLUTIONARY TREES

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 [11]. Poiché la vita sulla terra si estende dalle profondità del mare fino alla biosfera, si è reso necessario organizzarla in differenti ordini di grandezza. L'atomo è l'unità elementare che costituisce tutta la materia. Insiemi di atomi formano le molecole che, a loro volta combinandosi, formano le macromolecole. Insieme sono i costituenti delle cellule, le più piccole strutture classificabili come organismi viventi.
Ci sono quattro tipi di macromolecole essenziali per tutte le forme di vita:

- *Polisaccaridi*: macromolecole formate da aggregazioni di monosaccaridi, tra cui il fruttosio, il glucosio e così via. Sono riserve di energia pronta.
- *Proteine*: svolgono una vasta gamma di funzioni all'interno degli organismi viventi, permettendo le reazioni metaboliche, la replicazione del DNA, la risposta agli stimoli e così via.
- *Lipidi*: chiamati anche grassi, sono le riserve energetiche di deposito.
- *Acidi nucleici*: DNA e RNA, contengono e trasportano l'informazione genetica.

Le sezioni successive del capitolo sono dedicate al DNA, RNA e proteine.

1.1 DNA

Il *DNA* o *acido desossiribonucleico* è una macromolecola contenente il patrimonio genetico[1] degli esseri viventi [11], dunque ne detiene l'informazione ereditaria [7].
Porzioni specifiche di DNA contengono determinate informazioni, ad esempio il colore degli occhi, dei capelli e così via. Queste prendono il nome di *gene* [8].
Di seguito viene illustrata un'immagine del DNA, insieme ad una breve descrizione.

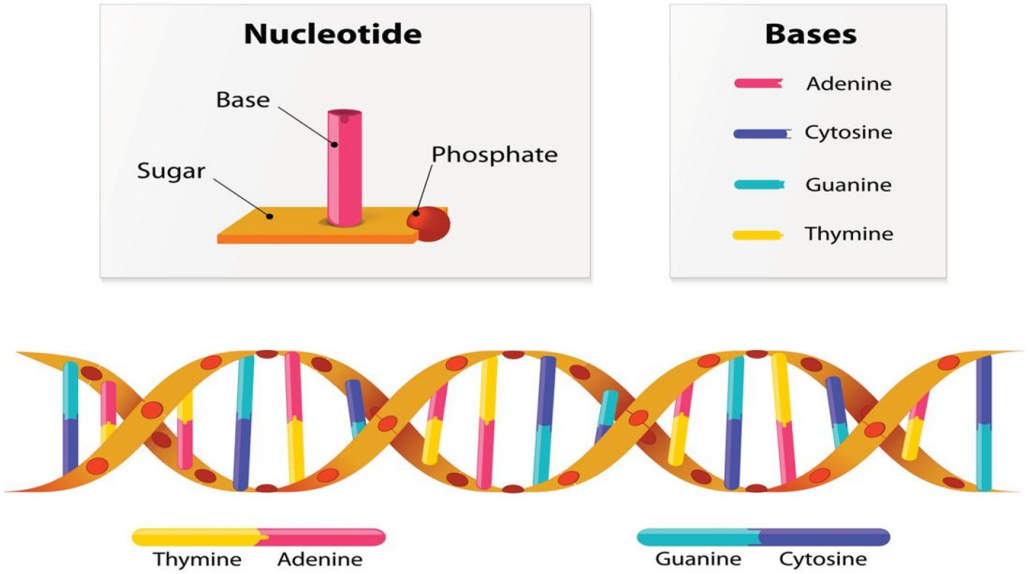

Figure 1: La struttura del DNA e del nucleotide.

La struttura è caratterizzata da una doppia elica di lunghezza variabile, dove ciascun filamento è formato da una sequenza di molecole chiamate *desossiribonucleotidi*.
Un desossiribonucleotide è composto da una molecola di zucchero, un gruppo fosfato ed una base azotata. Di quest'ultima, ne esistono quattro tipi:

- *adenina* (A);
- *timina* (T);
- *guanina* (G);

1 Il patrimonio genetico contiene tutte le informazioni genetiche di un organismo.

- *citosina* (C).

Una proprietà importante delle basi azotate è che sono biunivocamente legate tra loro: l'Adenina si può legare solo con la Timina (A-T), mentre la Guanina con la Citosina (G-C). Questo significa che i filamenti sono complementari e quindi se conosciamo la sequenza di basi di un filamento di DNA sappiamo anche la sequenza di quello complementare.

1.2 RNA

L'*RNA*, ovvero *acido ribonucleico*, è una macromolecola caratterizzata da una struttura a singolo filamento composta da una sequenza di lunghezza variabile di *ribonucleotidi*.
I ribonucleotidi si differenziano rispetto ai desossiribonucleotidi per una diversa molecola di zucchero e per la presenza della base azotata Uracile (U) che sostituisce la Timina.
Un tipo di RNA importante è l'*RNA messaggero* (mRNA), che trasporta l'informazione genetica contenuta nel DNA in una regione cellulare (citoplasma) in cui avviene la sintesi delle proteine[2].

1.3 PROTEINE

Le proteine sono le fondamenta di un organismo, infatti determinano la struttura e le funzioni delle cellule, ad esempio le cellule del cervello differiscono da quelle dei muscoli principalmente perché usano tipi diversi di proteine. La loro struttura è composta da sequenze di *aminoacidi* legati tra loro.
Sebbene esistano oltre cinquecento aminoacidi in natura, solo venti sono codificati dal codice genetico umano e pertanto utilizzati per la sintesi proteica.
La conversione dell'informazione genetica dal DNA in proteina, avviene in due processi, di seguito elencati:

1. *trascrizione*: viene prodotto l'RNA messaggero che trasporta l'informazione nel citoplasma dove avverrà la traduzione;

2. *traduzione*: l'informazione contenuta all'interno del mRNA viene convertita in proteine.

2 La sintesi proteica è il processo attraverso il quale vengono prodotte nuove proteine.

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, condividono gli stessi obiettivi:

- elaborazione ed estrazione delle informazioni;
- integrazione dei dati ottenuti tra sorgenti eterogenee;
- utilizzo trasparente ed efficiente dei dati in base al contesto scientifico, dalla raccolta, all'analisi, fino alla catalogazione;
- 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 [13] [23] [40].

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:

1. analisi del problema da affrontare;
2. raccolta ed analisi di dati statistici ottenuti a fronte di dati biologici in input;
3. creazione di modelli ed uso di strumenti matematici che possano essere applicati al problema in esame, al fine di sviluppare un algoritmo;
4. 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 [18]. 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, nel breve periodo, 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 avere una terapia individuale di maggiore efficacia, 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 nasce 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, di seguito elencate.

- *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 della frutta fino all'essere umano. Questa area di ricerca riguarda

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).

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 in questa area di ricerca risultano di particolare importanza gli *algoritmi di Clustering*, il cui obiettivo è quello di raggruppare i dati delle sequenze in insiemi (cluster) in modo veloce e preciso in base a determinati criteri, affinché gli elementi simili tra di loro siano nello stesso cluster mentre quelli differenti risiedono in altri. I

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

principali algoritmi di Clustering nella bioinformatica sono il *Clustering gerarchico* e l'*algoritmo k-Means*.

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*. 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 principalmente due. Prima di tutto, se l'espressione di un gene non conosciuto è simile a quella di un gene noto, è possibile ipotizzare che essi abbiano funzioni simili o che siano coinvolti nello stesso meccanismo biologico, portando a nuove scoperte scientifiche sui geni. 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 al 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 le entità attraverso la costruzione di alberi filogenetici. 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 tali alberi in modo estremamente accurato. Questi verranno spiegati esaustivamente nel capitolo successivo, in quanto oggetto principale della presente tesi.
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.

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

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

20 diversi tipi di aminoacidi e che se prendiamo una proteina composta da una sequenza di 70 di questi, è possibile ottenere $20^{70} (= 1.180591620717411303424 \times 10^{91})$ strutture diverse (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, nuovi dati biologici 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 basati sulla selezione naturale: 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 che riescono a risolvere problemi complessi velocemente, sfruttando la parallelizzazione automatica[9] ed il secondo che sono particolarmente ottimizzati per la ricerca genetica.
Da citare *Proteine Data Bank* [34], un vero e proprio archivio di acidi nucleici e proteine visualizzabili in 3-D.

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.

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

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.

- *Popolazione*: è un gruppo di organismi che vivono nello stesso luogo e che condividono determinate proprietà biologiche, pertanto sono della stessa specie.
- *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 con cui un allele genico si presenta in una popolazione.
  Preso in considerazione un gene in una popolazione, infatti, è possibile trovare alleli diversi con frequenze diverse.

La genetica delle popolazioni, quindi, si occupa di studiare la frequenza dei geni nelle popolazioni e la loro variazione nello spazio e nel tempo. Poiché in questa area di ricerca vengono coinvolti i processi evolutivi degli organismi, 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:

1. *popolazione iniziale*: data una popolazione con un determinato problema, viene scelto un gruppo di geni, rappresentati da 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 di fitness migliore, affinché trasmettano i propri geni alla generazione successiva;
4. *incrocio*: per ogni coppia di genitori viene scelto un punto di scambio tra i loro geni, di modo che i figli erediteranno informazioni genetiche mescolate (crossing over). Questa risulta essere la parte più importante di tali algoritmi;
5. *mutazione*: in alcuni casi ci possono essere delle mutazioni spontanee o indotte 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 di studio alla materia può essere bottom-up, partendo dai singoli geni fino all'organismo nella sua interezza, oppure viceversa, e quindi top-down.
Le applicazioni della bioinformatica in questa area di ricerca sono molteplici: l'ottenimento di dati ad alto throughput (high-throughput data) attraverso tecniche di analisi statistiche avanzate, e soprattutto la creazione di un linguaggio di Markup per i sistemi biologici, ovvero il *Systems Biology Markup Language* (SBML) [38]. 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) [36]. 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 informazioni non conosciute e potenzialmente utili a fronte di una grande mole di dati [18]. Con il passare del tempo e con il crescere della quantità di dati biologici, il data mining sta assumendo un ruolo sempre più 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 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*: l'eterogeneità delle informazioni biologiche ha reso necessario l'eliminazione dal set di dati da esaminare 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 proprio grazie 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 (dalla espressione genica al progetto Genoma Umano);
- modelli delle proteine in 3D (ad esempio il Proteine Data Bank [34]).

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, la loro modellazione è un punto cardine per la creazione di database biologici, nei quali è possibile individuare tre concetti fondamentali.

1. *Ordine delle sequenze*: le sequenze nucleotidiche e aminoacidiche possono essere modellate come entità statiche. Questo è dovuto principalmente al fatto che sono delle proprietà interne ad altre entità biologiche e che cambiano molto lentamente con il passare

del tempo (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 di DNA, si ottengono delle sequenze di aminoacidiche, grazie ai processi di trascrizione e traduzione.

3. *Relazioni spaziali tra le molecole*: come già detto nella sottosezione "bioinformatica strutturale", la struttura 3D delle macromolecole (in particolar modo le proteine) è alla base per la comprensione delle loro funzioni, pertanto, a livello di database, è necessario definire le relazioni tra le varie entità che compongono una macromolecola.

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 [20], National Center for Biotechnology Information (NCBI) [32] ed il già citato Proteine Data Bank [34].

3

CAPITOLO 3: ALBERO EVOLUTIVO

Una delle sfide più importanti della bioinformatica, nonché l'obiettivo principale della filogenetica, è la costruzione degli alberi evolutivi.
Ricordando la definizione di albero, ovvero un grafo non orientato connesso e aciclico [33], *L'albero evolutivo* o *albero filogenetico* è un diagramma che rappresenta le relazioni evolutive tra le varie entità biologiche [1] (animali, piante, virus e così via) [6]. La loro peculiarità consiste nel poter utilizzare dati diversi tra loro, genetici, genomici e morfologici.
Gli alberi evolutivi possono essere suddivisi in due categorie: alberi radicati e non radicati.

3.1 ALBERO RADICATO

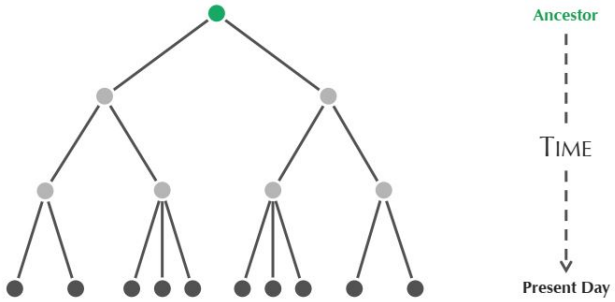

Figure 2: Esempio di albero radicato.

L'albero radicato o *albero con radice* è un albero in cui i nodi (o vertici) rappresentano le entità biologiche, mentre gli archi rappresentano la loro

1 I termini albero evolutivo, albero filogenetico e cladogramma sono spesso usati in modo intercambiabile per descrivere la stessa cosa, ovvero le relazioni evolutive tra entità biologiche. Questo perché l'uso del vocabolario non è sempre coerente nella letteratura scientifica, sebbene il contesto sia lo stesso [16].

evoluzione nel tempo (figura 2) [17]. Esso si sviluppa a partire da un nodo speciale, chiamato *radice* (il vertice verde nella figura) e si estende fino alle foglie. I vertici che hanno grado[2] maggiore di uno, definiti *nodi interni*, sono gli antenati, mentre quelli con grado esattamente uguale ad uno, definite *foglie*, sono le entità attualmente esistenti. La radice, quindi, è l'antenato comune a tutti i vertici dell'albero.
Lo scopo dell'albero non è solo quello di conoscere la radice, ma anche i legami tra le entità.

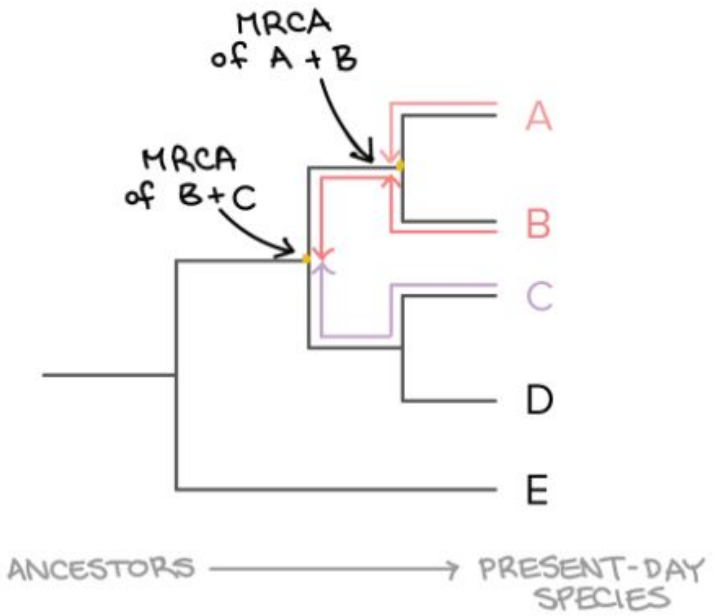

Figure 3: Esempio di albero radicato che mostra le relazioni tra le entità biologiche.

Come mostrato nella figura 3, le entità risultano più legati tra loro rispetto agli altri se hanno l'antenato più recente in comune, ad esempio A è più legato a B rispetto a C. La sigla "MRCA", infatti, indica il Most Recent Common Ancestor.
Nel caso in cui non è presente la radice, si parla di albero non radicato.

2 Il grado di un vertice v è dato dal numero degli archi incidenti su v [33].

3.2 ALBERO NON RADICATO

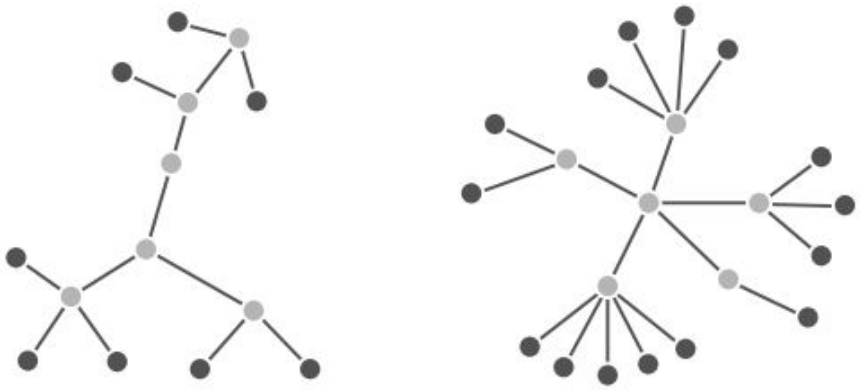

Figure 4: Esempi di alberi non radicati.

L'albero non radicato o *albero senza radice* è un albero in cui i nodi rappresentano le entità biologiche, mentre gli archi rappresentano la loro relazione, pertanto non richiedono la conoscenza della radice (figura 4) [17].
Perché usare gli alberi non radicati invece di quelli con radice? Le motivazioni sono varie.

- Gli alberi senza radice possono essere considerati una generalizzazione di quelli radicati. Questo consente agli scienziati di formulare ipotesi in merito alle relazioni tra le entità in modo più intuitivo.
- Molti degli algoritmi utilizzati costruiscono alberi non radicati e solo successivamente viene trovata la radice, se necessario.

3.3 METODI PER LA COSTRUZIONE DEGLI ALBERI EVOLUTIVI

Gli algoritmi utilizzati per la costruzione degli alberi evolutivi possono essere categorizzati in due metodologie:

- *Metodi basati sulla distanza*: vengono raccolti i dati in una matrice, definita matrice delle distanze. Nel capitolo successivo ne verranno mostrati gli algoritmi;
- *Metodi basati sui caratteri*: vengono usate le sequenze di DNA e di aminoacidi.

4

CAPITOLO 4: ALGORITMI BASATI SULLA DISTANZA

In questo capitolo vengono illustrati i principali algoritmi utilizzati per la costruzione degli alberi evolutivi. L'obiettivo è quello di trovare una soluzione al cosiddetto *problema degli alberi basati sulla distanza*, ma prima è necessario introdurre alcuni concetti, tra cui la matrice delle distanze.

4.1 MATRICE DELLE DISTANZE

Dati due punti, x e y, la *distanza* può essere vista come loro "lontananza" in uno spazio k-dimensionale. Nella fattispecie, è una funzione $d(x, y)$ che possiede le seguenti proprietà [29]:

1. *non negatività*:

$$d(x, y) \geqslant 0 \qquad \forall\, x, y \in R^k$$

2. *identità*:

$$d(x, y) = 0 \;\leftrightarrow\; x = y$$

3. *simmetria*:

$$d(x, y) = d(y, x) \qquad \forall\, x, y \in R^k$$

4. *disuguaglianza triangolare*:

$$d(x, y) \leqslant d(x, z) + d(y, z) \qquad \forall\, x, y, z \in R^k$$

Date n unità, calcolando la distanza per ogni coppia di elementi[1] si ottiene una *matrice delle distanze* $n \times n$, definita nel seguente modo [24]:

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & d_{2n} \\ d_{31} & d_{32} & 0 & \dots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & 0 \end{pmatrix} \qquad \text{dove } d(x_i, x_j) = d_{i,j}$$

Poiché è costruita a partire dalle distanze, ne eredita le proprietà precedentemente elencate.
Ciascun valore $d_{i,j}$ può assumere significati diversi in base al contesto, ad esempio, può indicare il numero di simboli diversi tra i geni i e j nell'allineamento di sequenze di DNA[2], come mostrato nell'esempio sottostante:

SPECIE	ALLINEAMENTO	MATRICE DELLE DISTANZE			
		Umano	Scimpanzé	Foca	Balena
Umano	ATGTAAGACT	0	3	7	5
Scimpanzé	ACGTAGGCCT	3	0	6	4
Foca	TCGAGAGCAC	7	6	0	2
Balena	TCGAAAGCAT	5	4	2	0

Figure 5: Esempio di matrice delle distanze.

Dalla figura 5 è possibile notare che la sequenza di DNA della foca risulta molto più simile a quella della balena, in quanto la distanza è 2, piuttosto che con l'umano, la cui distanza invece è 7.

1 Ci sono vari modi per calcolare la distanza tra una coppia di elementi, ad esempio attraverso la distanza Euclidea, quella di Manhattan, di Minkowski e così via.

2 'allineamento è il processo attraverso il quale si misura la similarità tra due o più sequenze.

4.2 PROBLEMA DEGLI ALBERI BASATI SULLA DISTANZA

Gli algoritmi basati sulla distanza utilizzano la matrice delle distanze per costruire gli alberi evolutivi, dove le foglie corrispondono alle entità biologiche presenti nella matrice, mentre i nodi interni rappresentano gli antenati non noti. Per poter conoscere quale è la distanza tra due foglie, e quindi conoscere quanto sono legate tra loro, è necessario associare un valore non negativo (peso) a ciascun arco, pertanto la lunghezza del cammino in tale albero è la somma dei suoi pesi. Si definisce, quindi, la *distanza evolutiva* tra due entità biologiche corrispondenti alle foglie i e j di un albero T come la lunghezza dell'unico cammino che li collega, ed è indicato come $d_{i,j}(T)$ [17]. In altre parole è dato dalla somma dei pesi degli archi che ci sono tra i e j.
Si dice che un albero T si *adatta* ad una matrice delle distanze D se per ogni coppia di foglie i e j si ha che $d_{i,j} = d_{i,j}(T)$, ovvero l'elemento nella riga i e colonna j è uguale alla lunghezza del cammino che le collega (distanza evolutiva), in tal caso la matrice viene definita *additiva*. Qualora invece non esista un albero che si adatti alla matrice, allora è *non additiva*.
Si riporta di seguito un esempio che mostra un albero che si adatta alla matrice delle distanze mostrata nella sezione precedente.

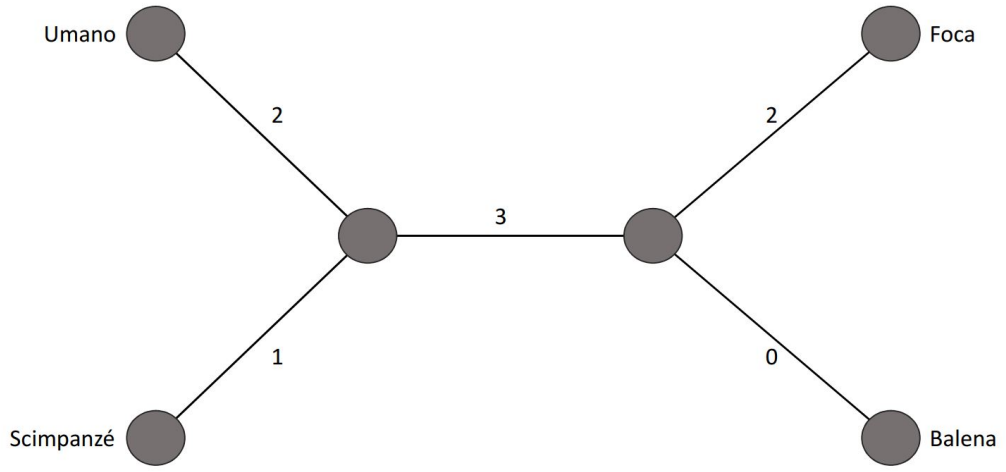

Figure 6: Albero evolutivo senza radice costruito a partire dalla matrice **adattiva** della figura 5.

Ci possono essere più alberi che si adattano ad una matrice, quindi come si può scegliere l'albero giusto? Si nota che quello in figura 6 ha tutti i vertici di grado diverso da due e viene definito *albero semplice*.

Una loro caratteristica importante è che per ogni matrice delle distanze adattiva esiste un unico albero semplice che si adatta alla matrice stessa. Adesso è possibile dare una definizione al problema accennato all'inizio del capitolo:

Problema degli alberi basati sulla distanza:
*Dato in **input** una matrice delle distanze adattiva si ottiene in **output** un albero evolutivo semplice.*

4.3 ALGORITMO PER IL PROBLEMA DEGLI ALBERI BASATI SULLA DISTANZA

L'obiettivo è quello di costruire un albero semplice T che si adatti alla matrice delle distanze D.
Si prenda in considerazione la matrice delle distanze mostrata nella figura 5:

$$D = \begin{array}{c} \text{specie} \\ u \\ s \\ f \\ b \end{array} \begin{array}{c} \begin{array}{cccc} u & s & f & b \end{array} \\ \begin{pmatrix} 0 & 3 & 7 & 5 \\ 3 & 0 & 6 & 4 \\ 7 & 6 & 0 & 2 \\ 5 & 4 & 2 & 0 \end{pmatrix} \end{array} u = \text{umano}, s = \text{scimpanzé}, f = \text{foca}, b = \text{balena}$$

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] Bear R., Herren C., Horne E., Rintoul D., Smith-Caldas M., Snyder B.
Building a phylogenetic tree.
https://www.khanacademy.org/science/biology/her/tree-of-life/a/building-an-evolutionary-tree (Cited on page 23.)

[7] Berg L., Martin W. D., Solomon E.
Biology.
Brooks Cole, 2010. (Cited on page 8.)

[8] Berk A., Darnell J., Kaiser A. C., Krieger M., Lodish H., Matsudaira P., Scott P. M., Zipursky L., .
Molecular Cell Biology.
W. H. Freeman (2008). (Cited on page 8.)

[9] 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.

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

[11] Cain M., Minorsky P., Reece J, Urry L., Wasserman S.
Campbell biology.
Pearson Higher Education, 2016. (Cited on pages 7 and 8.)

[12] 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

[13] Can T. (2013).
Introduction to Bioinformatics.
miRNomics: MicroRNA Biology and Computational Analysis.
DOI:
10.1007/978-1-62703-748-8_4 (Cited on page 11.)

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

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

[16] Choudhuri S.
Bioinformatics For Beginners.
Maryland, Elsevier Inc, 2014. (Cited on page 23.)

[17] Compeau P., Pevzner P.
Bioinformatics Algorithms, An Active Learning Approach, Vol II.
United States Of America, Active Learning Publishers, 2015. (Cited on pages 24, 25, and 29.)

[18] Dennis E. Shasha, Hannu Toivonen, Jason T. L. Wang, Mohammed J. Zaki.
Data Mining in Bioinformatics.
London, Springer, 2004. (Cited on pages 12 and 19.)

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

[20] Ensembl Website.
Ensembl.
`http://www.ensembl.org/` (Cited on page 21.)

[21] Gilbert D.
Phylogenetic trees.
University of Glasgow, Department of Computing Science.
`http://people.brunel.ac.uk/~csstdrg/courses/glasgow_courses/website_bioinformaticsHM/slides/phylo.pdf`

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

[23] ICAR CNR: Istituto Di Calcolo E Reti Ad Alte Prestazioni.
Bioinformatica.
`https://www.icar.cnr.it/bio-informatica/` (Cited on page 11.)

[24] Ingrassia S.
Statistica Aziendale 2.
http://www.ecostat.unical.it/Didattica/Statistica/didattica/StatAziendale2/StatAz2_cap1_2.pdf
Dispense, Università degli studi di Catania. (Cited on page 28.)

[25] 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

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

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

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

[29] Mola F.
Distanze ed indici di similarità.
http://people.unica.it/francescomola/files/2014/11/Cap-IV-_DISTANZE-ED-INDICI-DI-SIMILARITA.pdf
Dispense, 2014. (Cited on page 27.)

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

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

[32] National Center for Biotechnology Information Website.
NCBI.
https://www.ncbi.nlm.nih.gov/ (Cited on page 21.)

[33] Olivieri M., Verri M. C.
Algoritmi E Strutture Dati II.
Dispense, Anno Accademico 2013/2014. (Cited on pages 23 and 24.)

[34] Proteine Data Bank Website.
PDB.
http://www.rcsb.org/ (Cited on pages 17, 20, and 21.)

[35] 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

[36] Systems Biology Workbench Website.
SBW.
http://sbw.sourceforge.net/ (Cited on page 19.)

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

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

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

[40] Tramontano Anna (2003).
La grande scienza. Bioinformatica.
http://www.treccani.it/enciclopedia/la-grande-scienza-bioinformatica_%28Storia-della-Scienza%29/ (Cited on page 11.)