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: 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, ad esempio i dati raccolti in campo astronomico saranno di natura diversa rispetto quelli raccolti in campo biologico, condividono gli stessi obiettivi, come riportato nella pubblicazione **[?]** *X-Informatics: Practical Semantic Science*:

- Processamento ed estrazione delle informazioni
- Utilizzo trasparente ed efficiente dei dati in base al contesto scientifico, dalla raccolta, all'analisi fino alla catagolazione
- Integrazione di dati ottenuti tra sorgenti eterogenee
- Interazione con la raccolta dati adattata e personalizzata per l'utente
- 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.

1.1 CHE COSA È LA BIOINFORMATICA?

Non esiste un unico modo con il quale 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 nell'introduzione al capitolo, va aggiunto quello che risulta l'obiettivo principale di questa disciplina, ovvero quello di aumentare la conoscenza di tutti quei processi di natura biologica.
A prescindere dalla natura del problema da affrontare, è possibile individuare un approccio standard, suddiviso in cinque steps:

- Studio ed analisi del problema da affrontare
- Collezionamento ed analisi di dati statistici 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 genomica. Per **high-throughput data** si intendono 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.
Nel breve periodo, ad esempio grazie al *progetto genoma umano*[1], si possono scoprire nuovi geni legati alle malattie e nuovi bersagli molecolari, ovvero quei processi biologici, intesi come proteine, recettori, pathway biochimici etc su cui si può intervenire per modificare il decorso di una malattia.
Nel lungo periodo sarà possibile scoprire eventuali reazioni avverse ai farmaci da individuo ad individuo in base a dei tests, al punto tale che,

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 in esso i geni contenuti. Il progetto è cominciato nel 1990, per poi essere completato nel 2003 ed ulteriori ricerche sono ancora in corso.

grazie all'informazione genetica ottenuta attraverso strumenti strumenti informatici, sarà possibile personalizzare l'uso di farmaci, portando ad una migliore efficacia alla terapia individuale, riducendo o addirittura eliminando possibili effetti collaterali.

1.2 AREE DI RICERCA

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

1.2.1 *Analisi dei genomi*

Uno dei principali focus della bioinformatica riguarda l'analisi dei genomi[2] degli organismi il cui sequenziamento[3] è già stato completato, dal moscerino della frutta fino all'essere umano. L'analisi dei genomi è un'area di ricerca relativa non solo alla bioinformatica, ma anche alla genomica, ovvero quella disciplina che studia la struttura, il contenuto, la funzione e l'evoluzione del genoma. Ma perché analizzare i genomi? un gene viene sequenziato per conoscere la sua funzione ed eventualmente per modificarne la sua funzione. La conoscenza dell'intero genoma di un organismo fornisce le sequenze di tutti i suoi geni, permettendo così di identificare e manipolare i geni importanti che influenzano il metabolismo, la differenziazione e lo sviluppo cellulare e i processi patologici negli umani, animali e nelle piante.
L'obiettivo in questa area di ricerca per la bioinformatica è quello di identificare e alterare tutti quei geni che abbiano una particolare funzione biologica attraverso strumenti computazionali.

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

3 Il sequenziamento è un processo mediante il quale viene determinata la struttura primaria delle macromolecole (composizione atomica e legami), quali il DNA, RNA e proteine.

1.2.2 *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[4] 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 il nome di un gene oppure la fonte, dopodiché vengono salvate all'interno di database che prendono il nome di *database biologici* (vedere sottosezione 1.2.9).
L'analisi delle sequenze di DNA permette di venire a conoscenza dell'informazione genetica che viene trasportata all'interno di un suo segmento, grazie al quale, ad esempio, è possibile individuare i cambiamenti di un gene che possono causare una potenziale malattia.
Una volta estratto il DNA, vengono create migliaia e migliaia di copie di un singolo frammento, in quanto non è singolo frammento non risulterebbe sufficiente. Questi campioni vengono inseriti in dei macchinari chiamati *DNA Sequencer* che svolgono il compito di 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 clusters giocano un ruolo chiave all'interno della bioinformatica, infatti data la grande quantità di dati da gestire e manipolare, è possibile raggrupparli in insiemi (clusters, appunto) secondo determinati criteri, di modo che gli elementi che sono simili tra di loro stiano nello stesso cluster mentre quelli che sono differenti risiedono in altri clusters. I principali algoritmi di Clustering nella bioinformatica sono il *Clustering gerarchico* e l'*algoritmo k-Means*, che verranno spiegati successivamente.

1.2.3 *Analisi dell'espressione genica*

Prima di procedere con la definizione di "Analisi dell'espressione genica", è necessario fare una piccola introduzione. Quando un gene è attivo, esprime un codice genetico, definito *espressione genica*.Successivamente le

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

sequenze che formano tale codice genetico vengono copiate (trascrizione), producendo così RNA messaggero (mRNA), che a sua volta sarà coinvolto per la creazione delle proteine.
Con espressione genica, quindi, si intende la manifestazione fenotipica[5] di uno o più geni, in altre parole quando la sua (o la loro) informazione genetica viene trascritta e tradotta in una proteina.
L'analisi dell'espressione genica si occupa di studiare il quantitativo di RNA messaggero(mRNA) presente in un determinato gene. L'approccio ideale è, pertanto, trovare il gene, analizzarlo ed estrarre le informazioni fenotipiche interessate.
In questa area di ricerca il data mining e gli algoritmi di Clustering giocano un ruolo di fondamentale importanza, in quanto i primi consentono estrarre le informazioni mentre i secondi permettono di classificarle in gruppi. Come già riportato nella sezione "Analisi delle sequenze", gli algoritmi di Clustering "Clustering gerarchico" e "algoritmo K-means" risultano di fondamentale importanza.
Ma perché l'analisi dell'espressione genica è importante? Le motivazioni sono molteplici. Prima di tutto, se l'espressione genica di un gene non conosciuto è simile all'espressione genica di un gene conosciuto, è possibile che essi abbiano funzioni simili o che siano coinvolti nello stesso meccanismo. 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, infatti trova applicazione pratica nella lotta contro il tumore, predicendo eventuale metastasi.

1.2.4 *Filogenetica*

La filogenetica è quel campo della bioinformatica e della biologia evolutiva[6] che studia le relazioni evolutive tra vari gruppi di organismi. Tradizionalmente si basava sul confronto delle caratteristiche morfologiche degli organismi, oggi invece si usano i dati molecolari delle sequenze, permettendo la costruzione di alberi filogenetici con molta accuratezza.
L'albero evolutivo o *albero filogenetico* è un diagramma che rappresenta le relazioni evolutive tra i vari organismi(spiegato successivamente), la cui invenzione è dovuta a Charles Darwin, nel 1837. Una delle peculiarità è

5 Fenotipo: le caratteristiche manifeste di un organismo.
6 La biologia evolutiva si occupa dello studio delle origini ed evoluzioni delle specie.

che si possono costruire alberi 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 ricostruire ed analizzare gli alberi filogenetici.
Le applicazioni della filogenetica sono molteplici:

- 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, in cui hanno trovato impiego.
- Classificazione: fornisce dei metodi di classificazione degli organismi viventi in modo efficiente ed accurato.
- Forense: può essere usata per valutare delle prove di DNA in casi giudiziari.
- Identificazione dell'origine degli agenti patogeni: insieme all'analisi delle sequenze (sottosezione 1.2.2) è possibile studiare ancora più a fondo gli agenti patogeni[7], impedendo 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: fornisce informazioni aggiuntive per impedire la scomparsa di animali o vegetali, garantendone, quindi, la loro conservazione.

1.2.5 *Bioinformatica strutturale*

Le macromolecole (DNA, RNA e proteine) svolgono la loro funzione all'interno di un organismo grazie alla loro struttura nello spazio tridimensionale e quindi dell'avvolgersi delle loro sequenze di aminoacidi di cui sono composti. Tuttavia conoscere tale struttura non è affatto banale, ad esempio se prendiamo in considerazione tutti e 20 gli aminoacidi delle proteine ed una proteina composta da 70 aminoacidi, otteniamo $20^{70} (= 1.180591620717411303424 \times 10^{91})$ possibili strutture diverse che si

7 Gli agenti patogeni sono i virus,i batteri, etc...

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

1.2.6 *Genetica delle popolazioni*

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

8 Con predizione si intende predire la 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 (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 1.2.5).
È possibile dare una sguardo più approfondito a tali algoritmi, individuando cinque fasi fondamentali:

1. Popolazione iniziale: data una popolazione con un determinato problema, vengono scelti un gruppo di geni, rappresentate 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.

1.2.7 *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 connesse tra loro (midollo spinale, nervi, cervello e 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)[**?**]. 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)[**?**]. Il SBW è un insieme di strumenti che permettono di creare, visualizzare e simulare le reti tra le entità biologiche che compongono un sistema biologico.

1.2.8 *Data mining*

1.2.9 *Database biologici*

1.2.10 *Bioimmagini*

1.3 STORIA

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

[4] 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. (Cited on page 7.)

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

[6] 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

[7] Can T. (2013).
Introduction to Bioinformatics.

miRNomics: MicroRNA Biology and Computational Analysis.
DOI:
10.1007/978-1-62703-748-8_4

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

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

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

[11] 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

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

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

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

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

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

[17] Proteine Data Bank Website.
PDB.
http://www.rcsb.org/ (Cited on page 13.)

[18] 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

[19] Systems Biology Workbench Website.
SBW.
http://sbw.sourceforge.net/

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

[21] Systems Biology Markup Language.
SBML.
http://sbml.org/ (Cited on page 15.)

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

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