UNIVERSITÀ DI PISA

Corso di Laurea Triennale in Informatica (L-31)

TESI DI LAUREA

AlphaFold e le prospettive della bioinformatica

Relatore
Prof. **Paolo Milazzo**

Candidato
Ludovico **Venturi**

ANNO ACCADEMICO 2020/2021

Riassunto

Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe.
Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe. Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe.

Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe. Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe.
Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe.
Va posto al centro della seconda pagina e non dovrebbe superare le 20 righe.

Indice

Trasformare l'esperienza dell'università in qualcosa di positivo, di progressivo, che può alimentare il fuoco delle mie passioni Fai qualcosa di specifico, renditi esperto.

Guida il lettore da 0 ad Alphafold facendolo meravigliare davanti alla bellezza della bioinformatica, e della vita.

Medita e poi scrivi qui: non passare da fonti terze. Non perdere il flusso.
Tu stai scrivendo qualcosa per te, non per il mondo. Scrivi, poi confrontati. Se ti confronti è normale che ti vedi inferiore. Come puoi invece essere inferiore a te stesso?

Ciò che conta è fare, fare, fare, mettere in pratica.

Hai scelto tu di uscire dall'informatica. Hai paura di risultare ignorante in biologia? Hai paura di esserti immischiato in un campo a te esterno e di sembrare "capiscione"?
1. Non ne sa quasi nulla nessuno dei prof 2.Non interessa loro 3. ho Mario Pirchio a cui chiedere aiuto 4. voglio uscire dall'informatica pura. Non mi fido. Non mi interessa. Qui mi interesso 5.affronta la responsabilità. Ho la responsabilità di creare la mia strada e crederci, di laurearmi per mio padre e la mia famiglia.

Mentre disegnavo ho notato che ciò che mi spingeva a a migliorare il disegno era riuscire a intravedere il risultato finale in quello che stavo facendo. Non stavo tracciando una linea su un foglio. Stavo facendo piccoli passi per mettere su carta ciò che vedevo dentro di me (non nella mente, ma nel cuore).
Realizzavo una piccola parte di me al di fuori di me. E vedere che ciò che stavo creando si stava avvicinando a ciò che avevo in mente mi dava una soddisfazione immensa. E questa felicità mi spingeva tantissimo a continuare e a migliorarmi.
Voglio scrivere questo documento per realizzare una piccola parte di me all'esterno di me. L'obiettivo del disegno era realizzare un ritratto di Thich Nhat Hanh, per esprimere la mia gratitudine nei suoi confronti.

Obiettivo finale: realizzare un documento riguardante il background della bioinformatica e lo studio di AlphaFold per esprimere la mia speranza che l'informatica possa essere usata per il bene della Vita, che ci possa avvicinare ad una comprensione maggiore di essa e di quanto ogni fenomeno sia interrelato.

La tesi serve a dimostrare una ipotesi che avete elaborato dall'inizio, non a mostrare che voi sapete tutto

Ludo non dimenticarti quanta luce hai, sei ricco di una bellezza tanto speciale, non lasciare che altri te la nascondano
... ti ringrazio per ciò che sei...
Tutto ciò che sei, che dici, che fai è meraviglioso
Dovresti sentirti in colpa con te stesso se invece abbandonassi tutto e tornassi a casa per paura di sbagliare
Ludo non hai bisogno di me, nè di tua madre nè di nessun altro. Tu sei una persona davvero meravigliosa, sei forte, hai tanta luce in te. Una cosa che penso di sapere è che potresti fare qualsiasi cosa, andare da qualsiasi parte. E se lo vorrai io ci sarò in ogni caso, non hai bisogno del mio appoggio per raggiungere quello che vuoi, ma io sono qui, e ci resto per tutto il tempo che vorrai.
E potessi starti vicina ogni notte e risvegliarmi accanto a te la mattina farei il tifo per te direttamente dalla prima fila ;")
ci credo davvero nel risultato positivo che potrai scoprire tra un po', non demordere prima o poi arriverà esattamente quella cosa che stavi aspettando e tutto andrà a posto da sè [..Sophie..]

Una cosa per volta. Svuota il cervello. Adesso il mondo ti sembra pieno di problemi. Ne esiste solo uno per te: il tuo obiettivo. Se pensi a tutte le possibilità rimani fermo. Va solo in direzione dell'obiettivo. Poi al prossimo ci si penserà una volta raggiunto. Nessuno ti mette i bastoni fra le ruote, siamo con te. [papà]

Non solo hai dato voce alla tua vita, ma sei stato in grado di renderla poesia, sono veramente orgoglioso di te e di come ti ho rivisto dopo tanto tempo perché, te lo dico con tutto il cuore, hai fatto dei passi da gigante, dei passi enormi e veramente complimenti. [..diego]

Quando l'ansia bussa alla porta ringhio contro di lei: "ti affronto". Sono qui, avanti. Mi fermo e la guardo negli occhi. Affronto la vita, senza scappatoie.

Non abbiate paura di rischiare per non sbagliare. Mordete la vita. Sporcatevi le mani [Mattarella]

Introduzione

L'informatica: un potente strumento

Illustrare il mio obiettivo e la suddivisione del lavoro, dopo aver esposto la mia posizione sui rischi e le prospettive positive aperte dall'informatica.

Capitolo 1

Bioinformatica

1.1 Di cosa si occupa

Una parte importante della bioinformatica si occupa dell'utilizzo di strumenti informatici finalizzati a manipolare, archiviare e confrontare stringhe e sequenze di caratteri.
La bioinformatica tuttavia non si ferma all'analisi delle sequenze. Tra le più interessanti applicazioni bioinformatiche odierne vi sono quelle incentrate sull'analisi strutturale.
Difatti la bioinformatica pone le sue fondamenta nel campo della *structural bioinformatics*: per portare un esempio il database PDB (*Protein Data Bank*) nasce nel 1977 per archiviare coordinate atomiche e legami derivati dagli studi cristallografici sulle proteine [1].

1.2 Background filosofico

Buttaci un po' di filosofia della scienza e di quali cambiamenti potrebbe apportare alla struttura delle rivoluzioni scientifiche. Cita Fleck in qualche modo! Trova casi di cambi di paradigma e a "riscoperte" tornate alla ribalta grazie all'informatica. Magari l'informatica è un modo, analizzando tanti dati, di contrastare i bias nella scienza?

Capitolo 2

Proteine

Le proteine sono una classe di macromolecole con funzioni biologiche vitali permettendo il funzionamento di ogni sistema vivente. Riusciamo a pensare, parlare, a digerire il cibo, a muoverci grazie alle proteine.

Per questa ragione sono il target di grandi attività di ricerca e di applicazione biotecnologiche: dal combattere malattie infettive [2] al contrastare l'inquinamento ambientale [3].

Nel nostro corpo abbiamo un numero grandissimo di proteine: 10^{27}. Per usare una metafora di Ken Dill [4] potremmo dire che se si potesse ingrandire una molecola di proteina alla grandezza di un penny (diametro di 19mm [5]) il numero di protiene che una persona ha nel corpo è lo stesso del numero di penny che riempirebbero l'Oceano Pacifico.

Nonostante siano piccole, comparandole alle altre molecole del nostro corpo sono fra le più complesse e grandi.

–Foto proteina–

2.1 The central dogma of biology

2.2 Proteins and protein levels

2.3 Amino acids, nucleotides and codons

2.4 Protein structure characteristics

such as Domains, Motifs, Residues and Turns

2.5 Distograms

2.6 Phenotypes and genotypes

2.7 Protein Folding

È considerato uno dei problemi più imepgnativi degli ultimi 50 anni in biochimica. Utilizzando ancora le capacità divulgative di [4] si può immaginare una proteina come una collana composta da perle, dove ogni perla è un amminoacido e le perle possono avere 20 diversi colori.

Il punto del *protein folding problem* è capire come la stringa di amminoacidi codifichi la forma della proteina

Capitolo 3

Predizione della struttura delle proteine

L'analisi della struttura delle proteine è intrinsecamente complessa: "nessun'altra classe di molecole (piccole o grandi) esibisce una varietà di forme, dimensioni, struttura e movimenti comparabili alle proteine" [6].

3.1 Determinazione sperimentale della struttura delle proteine

Ci sono 3 tecniche sperimentali che possono essere usate per generare informazioni a risoluzione atomica sulla struttura delle proteine.

3.2 CASP

CASP (*Critical Assessment of Structure Predictions*) è una sfida biennale dove gruppi di ricerca si sfidano cercando di realizzare predizioni di strutture di proteine la cui sequenza amminoacidica è nota ma non la struttura determinata sperimentalmente, che verrà utilizzata per stabilire l'accuratezza dei metodi in gara.

Nel 2020 gli organizzatori del CASP14 hanno rinosciuto AlphaFold come soluzione del *protein–structure–prediction problem.*

3.2.1 Valutazione dell'accuratezza delle predizioni

The assessment of protein structure prediction tech- niques requires objective criteria to measure the similarity between a computational model and the experimentally determined reference structure. Conventional similarity measures based on a global super- position of carbon atoms are strongly influenced by domain motions and do not assess the accuracy of local atomic details in the model.

Le tecniche di valutazione della predizione della struttura delle proteine richiede criteri ogettivi sulla similarità tra un modello computazionale e la struttura di riferimento determinata sperimentalmente.
La misura di valutazione dell'accuratezza oggi è utilizzata è il lDDT (*local Distance Difference Test*) [7] Le misure convenzionali si basano su una superposizione globale di atomi di carbonio ma

3.3 Prima di AlphaFold

3.3.1 Machine Learning e biologia

Capitolo 4

AlphaFold

AlphaFold è un sistema di *Artificial Intelligence* (AI) sviluppato da DeepMind che realizza predizioni allo stato dell'arte sulla struttura delle proteine basandosi sulle loro sequenze amminoacidiche.

4.1 AlphaFold 1

4.2 AlphaFold 2

4.3 AlphaFold-Multimer e futuro

4.4 AlphaFold DB

4.5 Rischi per i metodi omologo

Rischi anemia falciforme (1 amminoacido diverso). Obiettivo tesi: come svicolare problemi dovuti a somiglianze sequenze ma funzione dverse. Spaventano!

4.6 Uso

4.7 Visualizzazione 3D del risultato

Capitolo 5

Sperimentazione di AlphaFold

5.1 proteina BFG-54g????

5.1.1 Confronto con altri metodi

Capitolo 6

Protein Engineering e campi applicativi

6.1 Enzymes engineering

6.2 Covid e Omicron

Conclusioni

considerazioni sulle porte aperte dalla bioinformatica Soddisfazione
Il problema del protein folding è risolto? (No).

Bibliografia

[1] F. C. Bernstein, T. F. Koetzle, G. J. Williams et al., "The protein data bank: A computer-based archival file for macromolecular structures," *Journal of Molecular Biology*, vol. 112, n. 3, pp. 535–542, 1977.

[2] M. Batool, B. Ahmad e S. Choi, "A structure-based drug discovery paradigm," *International journal of molecular sciences*, vol. 20, n. 11, p. 2783, 2019.

[3] B. C. Knott, E. Erickson, M. D. Allen et al., "Characterization and engineering of a two-enzyme system for plastics depolymerization," *Proceedings of the National Academy of Sciences*, vol. 117, n. 41, pp. 25 476–25 485, 2020.

[4] K. Dill. "The protein folding problem: a major conundrum of science: Ken Dill at TEDxSBU." [Online; accessed 6. Jan. 2022]. (ott. 2013), indirizzo: `https://www.youtube.com/watch?v=zm-3kovWpNQ`.

[5] "Cent (dollaro statunitense) - Wikipedia." [Online; accessed 6. Jan. 2022]. (dic. 2021), indirizzo: `https://it.wikipedia.org/w/index.php?title=Cent_(dollaro_statunitense)`.

[6] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.

[7] V. Mariani, M. Biasini, A. Barbato e T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, n. 21, pp. 2722–2728, 2013.