Biofisica Computazionale

Lorenzo Monacelli

6 febbraio 2015

Indice

Capitolo 1

Sequenziamento del DNA

Vedremo un paio di tecniche avanzate per ottenere le sequenze del DNA, le sequenze che si ottengono sono piccoli frammenti, bisogna ricostruire l'intera sequenza sulla base di questi frammenti. Gli algoritmi che si usano non sfruttano solo le sovrapposizioni. Spesso esistono algoritmi meno evidenti come logica ma più efficienti computazionalmente. Dal punto di vista della gerarchia in una cellula eucariote il DNA è compattato in forma di cromosoni: superavvolgimenti di DNA che a sua volta si arrotolano attorno a delle proteine dette *istoni*. Le basi azotate formano la doppia elica del DNA, arrotolata agli istoni e ancora compattata attorno al Cromosoma.

Ci sono delle funzioni per liberare delle parte di DNA che si vuole copiare nell'RNA. Il Dna viene ricopiato in RNA messaggero che viene a sua volta tradotto in una catena di aminoacidi che si avvolgono in una struttura tridimensionale per formare una proteina, tipicamente con l'aggiunta di qualche gruppo molecolare mediante enzimi. Tutto questo richiede un coordinamento di operazioni controllate. Se si vuole attivare una proteina si dovrebbe attivare tutto il meccanismo che apre il DNA, poi arrivano le proteine che permettono la trascrizione, e produrre le proteine che occorrono alla cellula.

Spesso invece le proteine sono già presenti e vengono modificate per attivarle, questo serve per velocizzare i processi. Ci sono varie modificazioni possibli: le più diffuse sono le fosforizzazioni. Poiché il fosforo è molto carico negativamente permette alla proteina di cambiare forma (portandola in uno stato di minima energia diversa). Qual è l'effetto di un aggiunta di un gruppo fosfato sulla struttura tridimensionale di una proteina? Nei genomi eucarioti la cosa è più complicata di così. La prima complicazioni è dovuta al fatto che i geni (le regioni che devono essere tradotte in qualcosa di funzionale) non sono contigue sul DNA, ma interrotta da regione che non servono a nulla (per formare la proteina che vogliamo ottenere). Queste regioni si chiamo *introni*. La molecola viene quindi prima trascritta in un RNA e poi tramite lo splicing si ottiene RNA messaggiero che elimina gli introni, che sono in numero e lunghezza variabie. Le regioni che vengono assemblate si chiamano esoni (Figura 1.1).

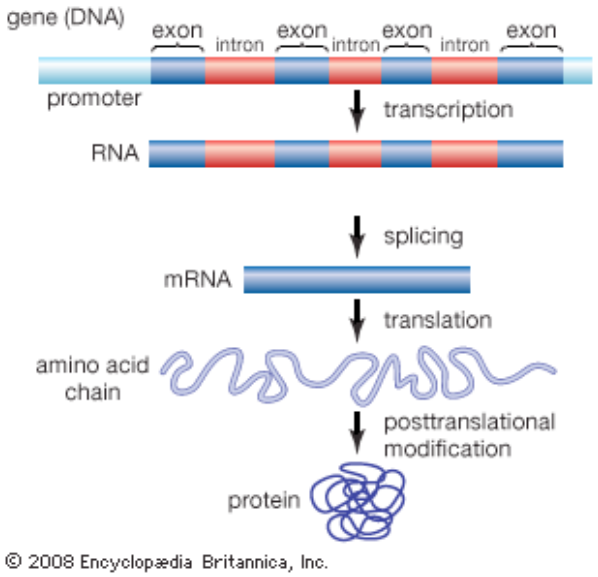


Figura 1.1: Come si passa dal DNA alla proteina.

Sprattutto negli organismi superiori esiste anche lo splicing alternativo, in cui gli esoni possono essere combinati in maniera diversa. Si può avere un RNA messaggero in cui un esone non è incluso, e così da una sola regione genonica si

possono ottenere proteine differenti. Questo processo normalmente porta alla formazione di proteine, ma non è detto che siano loro il prodotto fniale.

Le proteine svolgono il ruolo di:

- Trasporto e immagazzinamento
- Regolazione
- Protezione immunitaria
- Controllo di crescita e differenziazione
- Catalisi enzimatica
- Meccanismi di riparo

1.0.1 Gene

Il gene è l'unità ereditaria. Ha in se l'informazione di un carattere genetico, come ad esempio la risposta all'ingerimento dello zucchero, risposta all'aumento della temperatura, ecc.

È una regione degli acidi nucleici che non solo viene trascritta e tradotta, ma determina anche quando deve essere trascritta e tradotta. L'insieme di tutti i geni si chiama genoma. La bozza di tutto il genoma umano è stato concluso il 2003.

La sequenza del genoma è la seguenza di ATCG, una stringa. La regione codificante del gene inizia con la tripletta ATG, e termina con TAA, TAG o TGA. Ci sono introni ed esoni, il punto di separazione tra la fine di un esone e l'inizio di un introne è segnato dalle sequenze GT e AC. Ci sono delle altre regolette sulla regione di inizio e di fine, non sono particolarmente deterministiche. Quindi c'è il problema di trovare dove si trovano i geni. Ci sono molte poche regolarità negli introni o negli esoni. Il numero degli introni nei geni è distribuito in modo molto vario.

La dimensione dell'intero gene ha dimensioni da qualche centinaia di basi fino a decine di migliaia, e gli stessi introni hanno lunghezza variabile da qualche migliaia o qualche centinaia di basi. Nell'uomo sono circa tra i 20-25 mila geni. I geni non sono distribuiti uniformemente. ci sono anche regioni del genoma deserte. Isole ricche di geni e regioni molto più povere. C'è la storia che a sconvolto un po' tutti, una grande parte parte del genoma, quasi il 50 %, contiene DNA ripetitivo, che è generato da un meccanismo che ha a che vedere sulla copiatura. Noi non ci aspettiamo che abbia una grande rilevanza funzionale, abbiamo un 10 % ALU con regioni molto ripetute del DNA di cui non abbiamo capito il linguaggio. C'è solo l'1.5 % di genoma realmente codificante. Pian piano le regioni che sembrano non servire iniziano a diminuire, gli unici però analizzabili sono 1.5 % di tutto il codice genetico.

Il DNA si replica e l'informazione fluisce dal DNA -¿ RNA -¿ Proteina, al prodotto genico. Cosa abbiamo imparato guardando tutto questo? Come si è evoluto l'uomo se si è scambiato con altre specie tipo Neandertal o altre specie. Come facciamo a capire cosa c'è e dove? L'unico metodo che realmente abbiamo, oltre gli esperimenti sono metodi computazionali per cercare di avere un idea di quale può essere la funzione dei geni. La maggior parte delle analisi computazionale per ricavare informazione è quella di basarsi sull'evoluzione e sulla statistica.

L'uomo non sembra essere una specie molto stabile, immaginando che tutta l'evoluzione sia condenzata al Big Bang in un anno, il primo organismo arriva a fine febbraio (senza nucleo), per arrivare ad un organismo con un nucleo è giugno, i dinusauri sono arrivati in torno al 6 Dicembre e spariti al 18 dicembre. Noi siamo una specie giovanissima.

1.0.2 Evoluzione

Come funziona l'evoluzione? Gli organismi subiscono mutazioni casuali. Questa popolazione ha delle variazioni, ciascun membro della popolazione ha un DNA. Qualcuno è bravissimo a differenziarsi (virus) altri un po' meno (come noi). Se una particolare mutazione (rara), aumenta la fitness all'ambiente circostante, questo individuo avrà una maggiore probabilità di riprodursi. Si possono selezionare solo mutazioni che danno un vantaggio prima della riproduzione per la teoria evolutiva.

Mutazioni dannose che non hanno nessun effetto prima della riproduzione non contano, sono ammesse. L'anemia mediterranea è una mutazione di un gene dell'emoglobina che rende questa proteina insolubile e distrugge la forma de globuli rossi, se entrambe le coppie del gene dell'emoglobina sono strutturate il paziente muore. Poiché gli individui portatori sani appaiono svantaggiati questa mutazione dovrebbe essere stata cancellata dall'evoluzione. Tuttavia nella nostra zona è diffusa la malaria, che è sfavorita nel riprodursi nei geni non sani, e questo a favorito la mutazione malata dei geni.

Può capitare che due popolazioni indipendenti siano incapaci di incrociarsi nuovamente. Quando due sotto popolazione non possono dare origine a progenie fertine si chiamano specie diverse. Il genoma di queste specie però sarà più simile al genoma originario. Dal genoma di oggi possiamo ricostruire qual è l'albero della vita. Sappiamo che veniamo dallo stesso progenitore dello scimpanzé, ricostruendo questo meccanismo di speciazione.

Durante la seconda guerra mondiale c'era il problema seguente, si mandavano i caccia a bombardare, e la contraeria sparava a i caccia. Quali erano le regioni più importanti da proteggere per l'areoprano. Questo statistico ha mappato tutti i fori causati dalla contraeriea degli aerei bucati, e capito che occorre riforzare dove ci sono meno buchi (perché gli aerei che sono stati colpiti in quella zona non sono tornati).

Se nel DNA mappo le differenza tra topi e uomo le regioni molto conservate sono quelle più importanti, mentre quelle che variano molto saranno gli esoni (perché variazioni lì sono meno importanti).

Come si fa ad individuare la funzionalità di qualcosa? La si confronta con i suoi omologhi nell'evoluzione e si verifica se la omolgazione osservata è conservata o meno. Ho due elementi biologici, torvo l'algoritmo migliori per confrontarli e calcolare la *distanza*, poi qual è la distribuzione casuale dei valori e devo capire se il numero venuto fuori è compatibile con la distribuzione casuale o ha una bassa probabilità per avere una distribuione casuale.

Ad esempio posso confrontare l'emoglobina dell'uomo e quella del cavallo, che sono simili al 90 %, qual è la probabilità che questo sia dovuto a un caso? 10^{-500}, allora la ragione per cui sono dissimili è perché discendono dallo stesso gene ancestrale.

Due geni omologhi al 50 % significa che questi geni sono identici al 50% e quindi sono probabilmente omologhi. L'omologia è una proprietà binaria, due cose o sono omologhe o non lo sono.

Abbiamo lo stesso numero di geni del riccio di mare. Evidentemente l'aumento di complessità a partire da organismi meno evoluti a organismi più complessi non è dovuto ad un numero più elevato di geni, ma è dovuto soprattutto ai meccanismi di regolazione. Sono le interazioni che presumibilmente spiegano la gran parte della complessità. Questo ha protato nuovamente il discorso sulle parti di DNA fantoccio.

Nel replicamento del DNA avviene la mutazione dell'organismo. Dopo di che c'è una popolazione con una certa diversità.

Se c'è una grande diversità è molto maggiore la probabilità che qualche organismo sopravviva ad una catastrofe formando una nuova popolazione che avrà la proprietà di resistere a questa catastrofe.

Noi di diversità ce ne abbiamo molto poca. Le differenze tra ciascuno di noi è circa di 1 / 1000 lungo tutto il genoma. . Queste variazioni dove si trovano? Il 90 % delle variazioni presenti nell'uomo erano già presenti in Africa. Questo vuol dire che veniamo dall'Africa.

Se confrontiamo l'uomo con lo scimpanzé abbiamo una variazione ogni 100 nucleotidi. Nel topo, nonostante le sequenze rilevanti siano simili ci sono stati un sacco di riarrangiamenti. Il cromosoma 2 dell'uomo è sparso nei vari cromosomi. Nelle regioni importanti abbiamo circa il 10% di differenza. Dall'analisi genomica possiamo vedere come sono andate le cose. Confrontanto il genoma con quello del vermetto possiamo vedere quanto tempo è trascorso. Mentre la differenza tra due scimpanzè è 1/100 quella tra due umani è 1/1000. Noi siamo più vicini ad una catastrofe, ci siamo diversificati poco rispetto allo scimpanzè. Veniamo da un piccolo gruppo, che viveva in africa. Deve essere successo qualche cosa per cui è sopravissuto solo questo piccolo gruppetto. Sono una delle poche specie è sopravvissuta, cosa sia successo agli altri ancora non è chiaro.

1.1 Sequenziamento Del Genoma

Come si sequenzianoi genomi? Grazie ai virus e retrovirus, l'informazione può passare da RNA al DNA. Il virus retrotrascrivono l'RNA loro in DNA del nostro genoma. Il fatto che esiste un meccanismo che permette di retrocrascrive l'RNA in DNA. Significa che esistono delle proteine in grado di catalizzare questa reazione (ottengo la molecola di DNA dalla molecola di RNA). Questa scoperta non è affatto irrilevante.

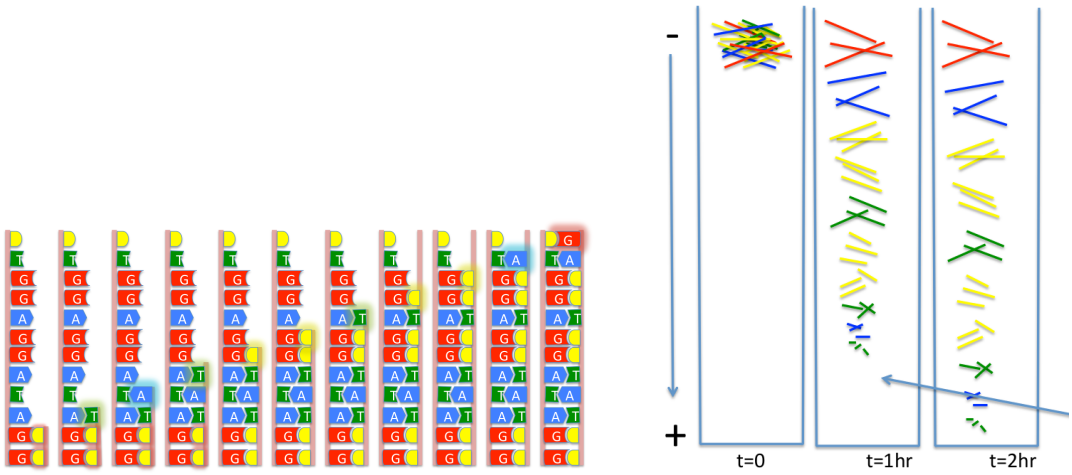
Ho una molecola di DNA che voglio replicare. Ho bisngno di una proteina che catalizzi questa reazione, che possa polimerizzare le ATGC per ottenere l'elica complementare. Il primo frammento di DNA a doppia elica è detto PRIMER, a cui la Polimerasi (proteina che replica il DNA) si attacca per continuare la trascrizione.

Occorrono polimerasi e primer, un elica si replica in una direzione e un'altra si replica nella direzione oppsote. Uno dei premi nobel sul DNA è stato dato per la reazione a catena della polimerasi (PCR), la PCR permette tutta l'analisi forenze, permette di amplificare il DNA esponenzialmente. SI prende una doppia elica di DNA e si sempara, e si aggiunge due PRIMER. Si mettono i nucleotidi e gli enzimi e ho due eliche, ripetendo questo procedimento con vari cicli posso avere tantissime molecole di DNA. Ci serve la polimerasi. Gli enzimi sono ottimali alla temperatura a cui funzionano. Se mettiamo a 90-80 gradi, si destrutturano e non funzionano più, ma d'altra parte queste due eliche di DNA devono essere separati a queste temperatura. Alzando la temperatura, si separano, rompo la polimerasi, poi la devo rimettere? Gary Murris ha scoperto che ci sono degli organismi che vivono a 80^o, la loro polimerasi è una polimerasi resistente alla temperatura, quindi io alzo e abbasso la temperatura spezzando e unendo le eliche. Si innesca la reazione di polimerizzazione, ne ottengo più. Siccome abbiamo scoperto che i virus hanno una proteina che è in grado di caratterizzare una reazione che retrotrascrive si può fare la PCR anche a partire dall'RNA. RNA non può essere identificare il genoma di nandertal perché non se ne trovano più tracce (poco stabile). Quindi la retrotraschrizione permette la replicazione dell'RNA.

Questo è molto interessante, prima di uscire dal nucleo all'RNA viene aggiunta una codina di poly.A (tante adenina). Nel nucleo ci sono tantissimi RNA, però, quelli che codificano, hanno una codina di POLYa. Se ho un supporto dove ho legato una coda di Poly T, e lo faccio interagire con la Cellula, solo l'RNA Poly A si legherà. In questo modo io

ottengo l'RNA che trascrive realmente. In questo modo posso sequenziare tutto l'RNA che trascrive in una cellula. Se prendo una cellula in un organo e una cellula di un tumore dello stesso organo, perché posso sequenziare tutto l'RNA delle due cellule e posso capire la differenza degli elementio funzionali di una cellula. Posso anche analizzare quali proteine vengono prodotte alla reazione di stimoli esterni. Possiamo ottenere specificatamente la sequenza di tutti gli RNA che catturo all'interno di una cellula. La quantità è irrilevante perché posso amplificarlo con la PCR.

Voglio vedre come si può sapere la precisa sequenza di ATGC. Il metodo classico è il metodo SANGER: prendo la doppia elica del DNA, un primer, invece di 4 nucleotidi ho sia i nucleotidi veri che altri modificati: sono nucleotidi che somigliano ai veri e vengono replicati. Questi terminano la catena e non permettono di legare nucleotidi successivi. Possiamo renderli fluorescienti.



(a) Basi Fluorescienti (b) Elettroforesi

Figura 1.2: Metodo di sequenziamento Sanger

QUesti sotituiscono le basi corrispondenti e fluorescono a colori diversi. Cosa succederà?

Casualmente nelle posizioni in cui c'è la A otterrò a volte la T che permette la prosecuzione della catena e a volta quella che interrompe la catena. Tutte le molecole che hanno avuto una T modificata verdi, A blu... però la lungehzza del frammento sintetizzato mi dice dove si è fermato e mi indicherà la posizione della base. Se si prendono tutti i frammenti mischiati si mettono in un gel, metto una differenza di potenziale, si muovono con velocità che dipende dalla loro lunghezza, vado a vedere il frammento più corto e posso automanticamente leggere i colori. Questo non è fatto a mano. Siamo in grado di ricostruire i colori e ricavare quali base erano quali. Il risultato viene fuori è una seguenze di colori che corrisponde alla seguenza di basi.

Con questo metodo si sequenzano fino a 800 basi di seguito (il genoma unamo è stato diviso in pezzetti da 800, sono 3 miliardi in totale XD). Il primo individuo si dice che è stato un uomo di Buffalo, non si è divulgato il nome di questo individuo. Su alcune cose sappiamo moltissimo. Analizzando l'inizio del gene per una specifica ci danno informazioni importantissimi.

1.2 Genoma umano

Il sequenziamento del genoma umano è stato completato da un consorsio pubblico nel 2003. Dal 2003 ad oggi sono stati sequenziati una serie di genomi umani, Gim Watzon si è sequenziato il suo genoma, e altri 1000 genomi di persone sane per vedere la variabilità all'interno della popolazione, ci sono enormi progetti per sequenziare vari tipi di cellule tumorali, e via discorrendo. Il desiderio di sequenziare tanti genomi è diventata parallela alla necessità di sequenziare genomi rapidamente: questa è chiamata Next-Generation Sequency, automatizzata con macchine in grado di produrre questi dati ad una velocità enorme. Questo ha creato un numero grande di problemi informatici. Sta diventando un problema per contenere tutti questi dati, se si hanno un enorme numero di dischi che generano un enorme calore che deve essere condizionata ad altissima energia. Il centro di caclolo di Cambrige è stato spostato a londra. Si sta pensando di costruire questi centri di calcolo nel nord europa. Come facciamo a sequenziare il DNA con tutta questa velocità? Riusciamo ad ottenere delle distribuzioni di lunghezza diverse. Possiamo avere dei frammenti dell'ordie di 300-400 nucleotidi[1]. Se noi prendiamo tutto l'RNA messaggero della cellula e otteniamo come output un file con tutte

[1]Ho bisogno di un Primer (un primo frammento sequenziato) e ci aggiungo una parte specifica, il cui primo pezzo è un adattatore. Se si replica l'elica e l'elica complementare viene fuori un campione amplificato Il passaggio successivo consente di denaturare la doppia elica, tutti i frammenti di dimensioni diverse, che conosco. Si prende un vetrino, in cui sono legati covalentemente dei frammenti comoplementari alla prima parte che abbiamo seguire la sequenza (l'adattatore), a questo punto aggiungo il DNA che si legherà nel suo primo frammento al

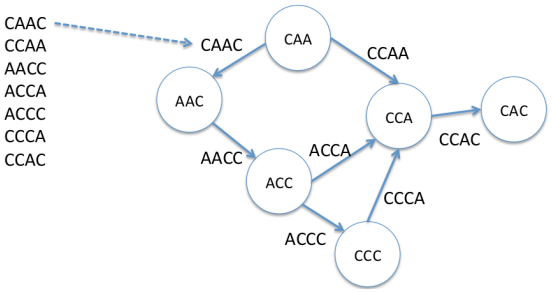


Figura 1.3: Grafo di Bruijn.

le sequenze dei frammentini con un valore di qualità di ciascuna base. Si può decidere se si hanno tanti dati di buttare tutti quelli di bassa qualità. Si può ricostruire quale era la sequenza di RNA mesasggero. Come si fa a ricostruire queste sequenze? Si utilizza la teoria dei grafi. Questo grafo particolare si chiamo grafo di *Bruijn*.

1.2.1 Teoria dei grafi

Questo grafo si costruisce mettendo per ogni sequenza si associa una connessione (edge). Ad esempio CAAC corrisponde alla connessione. Assegno ai nodi che lo precedono e seguono la connessione associo le basi N-1 (togliendo la prima o l'ultima). Se siamo riusciti a costruire un grafo che connette tutti questi nodi, quello che vogliamo è trovare all'interno di questo grafo un percorso che attraversi il grafo esattamente una volta. (Figura 1.3). Se ho un grafo, Si chiama percorso euleriano un percorso che visita ogni connessione esattamente una volta. Vogliamo un percorso euleriano. Eulero fu il primo a risolvere il problema della città di Konisberg (passare per tutti i ponti una volta sola).

Un percorso hamiltoniano è un percorso che visita ogni nodo esattamente una volta. Un ciclo euleriano parte e finisce sullo stesso nodo. Data una rete come si può trovare un percorso euleriano se esiste? Prendiamo le nostre sequenze e mettiamole in un nodo. Due nodi sono connessi se condividono k-1 basi. Se si collega il grafo passando una volta sola per le connessioni ho ricostruito la sequenza del gene. Un grafo è connesso se esiste un percorso che permette di andare da ogni vertice ad ogni altro vertice. Se il grafo è connesso, e ci sono al massimo o zero o due nodi con connessione dispari allora c'è il percorso euleriano. Come si fa a trovare il percorso euleriano quando esiste? (Vedo quanti sono i nodi con connessioni dispari, se sono o zero o due allora esiste).

L'algoritmo è il seguente: si parte da un nodo e poi si avanza ricorsivamente

- Se il nodo non ha primi vicini aggiungiamo il nodo al percorso
- Se il nodo ha un vicino, mantieni una lista dei vicini e procedi fino a quando non ci sono più vicini
- Per ogni vicino rimuovi la connessione tra il nodo di partenza e i suoi vicini
- Dopo aver processato tutti i vicini aggiungi il nodo al percorso

1.2.2 Confronti tra campioni

Se trovo un percorso euleriano ho trovato un modo di mettere assieme questi frammenti di sequernze. Tuttavia ci possono essere possibili percorsi alternativi, e quindi le sequenze potrebbero non essere esettamente uniche. Quando ci sono percorsi possibili si può valutare quali delle due alternative ha un supporto sperimentale maggiore (con maggiore qualità). Ci sono degli errori dovuti al sistema biolgico per cominciare (modifiche del RNA). Ci sono errori dovuti alla lettura ottica. Ci sono errori sperimentali nella costruzione dei supporti. E ci sono errori nella ricostruzione degli algoritmi euleriani. In realtà quello che abbiamo è sia una sequenza di un elica che quella della sequenza complementare, questa cosa può essere sfruttata per migliorare la nostra precisione. Si prende il genoma di un organismo si frammenta e si sequenzia. Un genoma di un batterio si può fare mentre per un organismo molto grande è complicatissimo, ci sono molte ambiguità. Ecco perché è stato importante il primo genoma umano, che fuge da riferimento per la ricostruzione di tutti gli altri genomi. Possiamo utilizzare praticamente di Routine per confrontare due trascrittomi. Se si ha una cellula di fegato e una cellula tumorale di fegato, cosa si è attivato per rendera incapace di controllarsi? Si può

frammento complementare del vetrino, quello che succede è che in quelle vicinanze ci sarà anche la sequenza complementare all'adattatore, il DNA formerà un arco ed andrà a legarsi al frammento complementare. Se aggiungo le quattro basi in DNA si replicherà. In ogni zona in cui è cascato un frammento di DNA c saranno tante copie di quel frammento. Questa replicazione si fa mettendo prima tutte le A con una certa fluorescenza, poi una T poi una C e una G. Quello che succederà è che in ciascuna posizione del vetrino se la prima base è una a vedrò una macchia blu una machia verde. All'inizio aggiungo una A e vedrò una fluorescenza caratteristica nella zona. Riesco a sequenziare milioni di frammenti in parallelo. Potrò ricostruire la sequenza di ciascuno di questi frammenti. Avremo una frequenza di DNA, con un identificativo della sequenza, e una riga che vi dice la qualità con cui quella base, che ci dice quanto bene quella macchia sia già stata sequenziata. Ci troviamo in questo file milioni di frammentini con l'errore su quella base. Adesso bisogna ricostruire quello che c'era in qeui campione

estrarre l'RNA, ricostruire i trascrittometri, e raccogliere delle differenze quantitative e qualitative. Ovviamente c'è un problema di significatività. C'è una variabilità biologica oltre i problemi sperimentali, c'è bisogno di una valutazione statistica accurata per affidare significatività. Il modo migliore per valutare la differenza statistica è quella di avere più replicati di ciascun esperimento. Un po' questi esperimenti sono abbastanza costosi quindi non si fanno prove ripetute, e spesso non è facile. La conoscenza dell'RNA messaggero dell'uomo ci aiuta a ricostruire quello che troviamo, sia per capire se ci sono differenze significative fra due individui. Si possono sequenziare i trascrittometri per mappare le variazione del genoma umano. Si prende una popolazione di paziente di una patologia, quella degli individui sani, e vedo come sono differenti. Ci sono patologie a singola variazione che sono state identificate. Ci sono patologie associate a cmbinazioni di patologie, e questo è difficilissimo da capire a causa dell'altissima variabilità del DNA tra gli individui.

Oltre a conoscere l'RNA messaggero si deve cercare di valutare una sequenza che mappa, cioè quanto sono condivise le informazioni tra gli esoni. Quantificare quanto RNA messaggero c'è è facile, se invece il gene ha più foprme finali occorre trovare un modo per quantificare l'abbondanza relativa.

Le banche dati sono depositati tutte le sequenze contenute con questi metodi, per questo scopo ci interessano tre banche dati: Gene Ontology che assegna ad ogni gene una funzione molecolare a tre livelli. Riconosce corpi estranei, trasporta ossigeno, fa parte del sistema di coagulazione del sangue. Il processo biolgico è la coagulazione del sangue e la componente cellulare. Per ogni gene è associato. KEGG è un database che conserva tutte le vie metaboliche. L'enziama 1 trasforma la molecola tizia in caio. C'è l'attività enzimatica di tutte le protenine. INTERPRO raggruppa le famiglie delle proteine omologhe, cioè discendono da uno stesso progenitore. Questi geni che sono più abbontanti in un campione piuttosto che un altro fanno parte di una certa cateogria (funzionalità?). Come si fa a sapere ciò? Si utilizza un calcolo statistico, la distribuzione pergeometrica. Ad esempio 377 geni sono più espressi nel campione 1. Assegnamo la loro funzione molecolare. Trovo un gran numero di geni che formano dellle Proteasi. Ne trovo 100. È significativa questa cosa? Se estraessi a caso 377 geni quanti ne troverei che fanno proteasi? Per fare ciò si usa una distribuzione ipergeometrica. Un modo per valutare che una certa sottopopolazione sia sottorappresentata o sovrarappresentata rispetto a quello che ci si rappresenterebbe a caso.

Un altro tipo di esperimento che si può fare è il fatto che al DNA ci sono delle proeine si possono legare al DNA per attivare o disattivare determinate geni. Se ho un fattore di trascrizione voglio sapere quali geni attiva e a quali valori si lega. Posso prendere il genoma, e in certe posizione ci sono legate vari fattori di trascrizione del fegato che permettono ai geni principali a quella funzione di essere trascritti. Aggiungo agenti chimici che legano permanentemente i fattori di trascrizioni al DNA. Poi aggiungo agli anticorpi contro i fattori di trascrizioni che vi interessa. Si frammenta il DNA in sequenze piccole, si mettono gli anticorpi su un supporto rigido.

Dobbiamo trovare i geni che ci interessano. Dobbiamo distinguere alcune caratteristiche di genomi Procarioti e Eucarioti. I batteri hanno necessità di replicarsi velocemente. Gli Eucarioti hanno una bassa densità genica, c'è il problema dello Splicing, bisogna identificare i vari pezzi di un gene. Nei genomi eucariata l'RNA messaggero ha una coda di poliA.

Come è fatto un genoma eucariota, ha una regione prima dell'inizio della trasuzione dove si legano ri fattori di trascrizione. AUG è il codone d'inizio, Introni ed Esoni, fino alla fine dove c'è il segnale Poli-A. C'è una regione 5primo non trascritto e una regione che si chiama 3 primo non trascritto. Codice di interuzione terminatore. Con gli introni di mezzo questa cosa non si può fare, occorre ricorrere a sistemi più complessi. Possiamo ricercare dei segnali di seuqenza. Possiamo costruire un modello statistico del gene (Modelli nascosti di Markof). Si possono sequenziare gli RNA messaggeri. Se siamo fortunati, e gli intrnoni non sono molo grandi e esoni molto piccoli, si possono sperare di trovare sul genoma. Oppure si può sfruttare il fatto che durante l'evoluzione ci sono evoluzioni ma geni dello stesso progenitore comune, posso vedere quali sono le regioni più conservate e vedere se queste corrispondono a dei geni. Per valutare se una certa popolazione e sotto o soprarappresentata in un campione si usa la distribuzione ipergeometrica:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Dove misuriamo la probabilità di trovare k erventi in n estrazione da una popolazione di N casi che contiene K eventi.

1.2.3 Matrici di sito specifico

Supponiamo di aver trovato sequenze simili ma non identiche, possiamo usare un metodo di matrici di posizione specifica per confrontarle. Prendiamo tutte le sequenze riconosciute da un certo fattore di trascrizione, e poi facciamo una tabella di conteggi. (Matrici posizione specifiche). Se il numero di frequenze che ho analizzato è molto alto queste approssimano la probabilità. (i numeri sono le posizioni nel gene). Abbiamo un problema, dire che non abbiamo mai osservato una A in prima posizione è una cosa, ma affermare che la probabilità sia nulla è un assersione forte. Aggiungo un valore piccolo a piacere a ciascuna cella del conteggio, in modo da tener conto di quella base. In modo da poter avere una tabella di frequenze che possa essere aiutato. Il fatto di eliminare gli zeri nella tabella è essenziale

per l'algoritmo. Si passa ai logaritmi dei numeri (essenziale non avere zeri), per poter fare sommare le probabilità invece di moltiplicarle. Dopo di chè, prendiamo la matrice di probabilità e voglio cercare altre sequenze sul genoma che possa essere riconoscuita. Faccio una finestra scorrevole. Parto dalla posizione 1 della mia sequenza, posso associare la probabilità che quella regione sia riconosciuta dal fattore di trascrizione. In questo modo ho un grafico che mi da la pobabilità che il fattore di trascrizione si leghi alla posizione centrata dalla mia tabella. Devo calcolarmi qual è la probabilità che ci sia una metching casuale. Questo può essere fatto ad esempio rimescolando le basi, generando 1000 sequenze fittizzie, rifaccio la stessa operazione e vedo qual è la media, e posso riconoscere se ci sono o meno significatività.

La matrice è così composta: Ciascuna riga corrisponde alle possibili basi (4 righe), e ciascuna colonna alla posizione. Gli elementi della matrice sono i logaritmi delle frequenze con cui è capitata quella base in quella determinata posizione quando nel codice genetico si è legato il fattore di trascrizione che ci interessa studiare. Data quindi una sequenza di DNA si fa scorrere questa tabella su tutte le posizioni sommando i valori ottenuti (ad esempio la sequenza è ACTGCT e io sommo il valore di A nella prima colonna, C della seconda,..., ottenendo una stima della probabilità che questa parte di codice genetico corrisponda realmente ad una zona in cui si legherà il fattore di trascrizione).

L'utlizzo della finestra scorrevole e rimescolare per caso valutando quello che mi aspetterei a caso per fare una differenza sono metodologie ricorrenti.

Posso anche calcolare qual è la quantità di informazione nelle sequenze di DNA, calcolando l'entropia di Shannon delle mie basi. In pratica ci dice quanta informazione c'è in quella posizione. L'entropia per la base a e posizione i, con frequenza $f_{a,i}$ è:

$$H_i = -\sum f_{a,i} \log_2 f_{a,i}$$

E il relativo contenuto di informazione è dato da:

$$R_i = 2 - (H_i + e_n)$$

Dove e_n è una correzione alla formula per tener conto che osserviamo un sottoinseme di dati e non tutti i dati. Per ogni posizione ottengo un valore dell'entropia di Shannon. Un modo per visualizzare in modo intuitivo è la matrice sitospecifica, è quella di fare un grafico in cui sull'asse y c'è il contenuto di informazione e sull'asse x la posizione. Poi si scrive una lettera che è tanto più alta quanto maggiore è il suo contenuto di informazione.

Esistono modi più computazionali, ci danno tante sequenze a monte di geni controllate da fattori di trascrizione, ci sono delle insiemi comune (MEME è l'algoritmo usato).

Abbiamo cominciato a vedere come analizziamo insiemi di sequenze che hanno una stessa proprietà, come si ricava un motivo di sequenza, dato un insieme di regioni genomiche che hanno una funzione analoga, per cercare di identificare le regolarità che identificano tutto questo, e possono essere utilizzate per trovare nuove ricorrenze di queste regolarità.

Bisogna estrarre il maggior numero possibile di informazioni che accomunano i dati sperimentali.

1.3 Identificazione di sottosequenze

Data una regione di DNA vogliamo sapere se vi è proesente qualche sottosequenza. Per ricercarlo possiamo prendere una finestrea scorrevole. Calcoliamo la probabilità che questa regione appartenga alla cella che vogliamo vedere. Calcoliamo che cosa ci aspettiamo dalla distribuzione casuale per indivisuare statisticamente quali regioni effettivamente appartengono ad una data sottosequenza. Anche in questo caso facciamo uso della matrice sitospecifica, con valori che approssimano la mia base.

Vogliamo identificare qual è la probabilità di ottenere una sequenza di un certo codice. Partendo dalla prima posizione avanziamo carattere per carattere calcolando il punteggio della matrice sito specifica della sottosequenza voluta.

Questo può essere utile per identificare geni che si comportano allo stesso modo, e quindi inferire che quei geni sono sotto lo stesso controllo, per esempio sotto una stessa proteina che permette loro di reagire allo stesso modo a stimoli estrni. Ho la sequenza di questi geni, sappiamo che queste proteine che controllano la trascrizione dei geni avviene prima dell'inizio dei geni. Ci sono caratteristiche comuni in questi geni? Ci sono delle sottosequenze in comuni a tutte questi geni?

1.3.1 MEME

Ci sono vari metodi, il MEME viene molto usato. Ho queste sequenze, che sono legate dalla stessa proteina, ad esempio le giunsioni introne – esone, e voglio vedere se ci sono sottosequenze molto simili. Mettiamo che queste sequznze siano proprio le stesse. Sono uguali dappertuitto tranne che per una base che non matcha bene. Se sapessi qual è questa regione potrei vedere se in altri geni è presente la stessa sottosequenza e inferire se sono anche loro parte del gruppo di geni.

Potrei confrontare tutte le sottosequenze con tutte le sottosequenze e me ne esco fra 3000 anni. Esiste un algoritmo che date delle sequenze, trova se ci sono delle sotto-sequezne comuni a delle sottosequenze? Se le mie sequenze le chiamo

s1 e sp, voglio trovare una siottosequenza della mia di posizione. Vorrei calcolare una similarità fra tutte le possibile sottosequenze di lunghezza data e prendere quella che minimizza la distanza.

Devo definire la distanza fra due sequenze, il modo più banale è quello di trovare la differenza tra due sequenze. La distanza di Humming è il numero di posizioni in cui le sequenze sono diverse. Io voglio trovare sottosequenze che abbiano la distanza più bassa possibile. Che minimizzino la distanza.

$$\text{argmin} \sum_{i<j} dist(s_i, s_j)$$

Posso confrontare tutte le sottosequenze ma è troppo lungo. Il metodo possibile è quello che si chiama MEME.

Supponiamo di avere 5 stringhe di diversa lunghezza. Si deve partire con 5 stringhe di dimensione del motivo che cerco. Si estraggnono queste sequenze casualmente dall'interno delle mie stringhe. Si esclude la prima stringa e si calcola la matrice sitospecifica dalle altre quattro e cerco tutte le sottosequenze, identificando i miglori match. Si ripete questo metodo fino a farlo convergere. Non c'è una soluzione formale che ci da che questo porta ad una soluzione reale, ma certamente ci sono più soluzioni.

Si parte da una desrizione random di un motovo, occorre avere un valore di Background (che mi aspetto per caso). Per ciascuna sequenza selezione una posizione casuale di partienza e la cerco dappertutto. In questo modo trovo dei motivi nelle mie sottosequenze. Questo è uno dei possibili modi di individuare sottosequenze simili che hanno alta probabilità di essere lì non per caso, e il metodo mi restituisce sottosequenze di lungehzza fra le varie regioni che ho analizzato. Posso prendere le sequenze di tutti i geni che mi nìinteressano, e chiedermi se ci sono sottosequenze comunindi questi geni che possono essere le regioni riconosciute da qualcosa che attiva o disattiva le regioni gniche.

Con 6 miliardi di basi del genoma unano, come si fa in questo marasma di ATGC a trovare le regioni che corrispondono al gene? Possiamo confrontare le regioni tra noi e il topo, perché le regioni più conservate sono presumibilmente più funzionale, e a regioni che sono variate tantissimo che è difficile immagonare che ci siano funzionalità.

Una volta che si è trovatra una sottostringa che permette di infierire che in quella regione si lega un fattore che ha a che vedere con un gene vado a vedere se c'è realmente un gene. Se questo risulta essere un buon metodo (specifico e sensibile) si fa scorrere la finestra su tutto il genoma, e se ci sono le regioni che contengono una sottosequenza che effettivamente permette il legame di una proteina, poiché queste si trovano a monte del gene posso inferire dove iniziano e terminano i geni.

1.3.2 Modelli di Markov

L'altro modo è quello di costruire un modello statistico del gene o di un altra regione. Nel DNA cisono regioni ricche di CG, dette isole CpG (coppia di C-G sulle due eliche complementari del DNA).

Una classe di mutazioni molto diffusa è la mutazione di C→T, questo fenomeno è molto meno frequente se la C è seguita da una G. Siccome ci aspettiamo che evolutivamente i geni siano protetti dalle mutazioni, in generale nelle regioni importanti le C sono spesso seguite dalla G. Le regioni in cui C non è seguita dalla G la mutazione di C in T porta ad una morte dell'individuo. Gli individui che sopravvivono sono quelle in cui la C è seguita da una G che preserva questa mutazione. Queste regioni possono essere di centinaia ma anche di migliaia di basi. Dove sono le isole CpG? Le si possono cercare i geni.

Una delle possibilità per identificare queste isole sono i modelli di Markov. Il processo di Markov è un processo in cui l'evento al tempo t dipende da quello che è successo precedentemente. Questi si chiamano gli eventi di Markov, se dipende dall'evento precedente è di ordine 1, se da due eventi precedenti è di ordine 2 ecc. Il caso delle isole CpG non mi basta vedere la frequenza della G , ma voglio vedere qual è la probabilità di avere G dopo C. Ho un insieme di stati presi da un alfabeto (ATGC) ho una certa probabilità di osservare un evento in una certa posizione. Avrò na certa probabilità di avere una transizione da C a G e un altra probabilità di avere una G data che la base precedente sia A, T, C, G.

Si prendono isole CpG e si calcolano tutte le possibili combinazioni di sequenze e si assume che questo sia una probabilità.

Abbiamo tutte le possibili coppie con dei valori di probabilità. Questo mi da la probabilità di osservare la c dopo G A o altro.

Quindi si elabora un altro modello di Markov per una parte del genoma che non è un isola CpG.

Ovviamente occorre ricordarsi di tenersi un test set per verificare la bontà del modello.

A questo punto si prende la sequenza che abbiamo e la si passa sul modello di Markof e si calcola la probabiltiò che la sequenza appartenga al modello CpG o al modello Non CpG.

Se i dati noti sono pochi ci sono vari trucchi per usare ugualmente il Test set (Live1Out, mi ricalcolo i parametri con tutte le sequenze meno 1, ogni volta escludendo una sequenza e provo il modello su quella sequenza, questo lo ripeto su tutte le sequenze, calcolo media e deviazione standard). Il punteggio di Markof può essere definito come:

$$S(x) = \ln\left(\frac{P(x|M+)}{P(x|M-)}\right) = \sum_{i=1}^{L} \ln\left(\frac{a^{+}_{x_{i-1},x_i}}{a^{-}_{x_{i-1},x_i}}\right)$$

Se S è positivo allora è più probabile che mi trovo in un isola CpG.

Posso quindi usare una finestra scorrevole, con il modello di Marcov al posto delle matrici sitospecifiche e vedere il punteggio del modello in funzione della posizione, prendere il massimo del punteggio e assegnare a quella regione il centro dell'isola CpG.

Si può complicare un po' il modello. Costruiamo un unico modello di markof che contenga sia le non isole CpG che le isole CpG.

Fino ad ora abbiamo fatto due modelli di Markof, uno positivo e uno negativo. Si possono unire i due modelli.

Possiamo calcolare all'interno di un modello tutte le probabilità. Data una A in una posizione, qual è la probiabiltà di avere prima una A T C P nel modello negativo e una ATCP nel modello positivo, se questa A appartiene a ciascuno dei due modelli? Una volta ottenute tutte le possibili probabilità di transizione si stima il percorso più probabile.

Non mi oposso permettere di calcolare la probabilità di tutti i percorsi: sono troppi. Tuttavia abbiamo ho un metodo per selezionare le migliori probabilità, Il concetto è la programmazione dinamica, se si vuole trovare il percorso più breve tra Roma e Milano, se il percorso migliore passa per Firenze, allora il percorso Roma-Firenze dovrà essere il miglior percorso possibile.

Quindi il percorso può essere ottimale solo se localmente è ottimale. Se voglio trovare il percorso ottimale all'interno di un modello nascosto di Markov. Trovare il percorso più probabile cercando qualunque sottopercorso che debba essere un percorso ottimale fino a quel punto. Occorre decidere quali siano le cose positive negative o frecce utilizzando il trining Dataset, Quindi calcolerò tutte le possibili transizioni, facciamo un modello. Un modello di markof può essere complicato all'infinito. Se riusciamo ad avere dataset sufficientemente ampio si possono costruire i modelli di markof riguardo a tutto. Come si trovano i geni veramente, potremo trovare le isole CpG, i promotori gli ATC e altre cose. Oppure costruiamo un unico modello di Markopf nascosto che includa tutto ciò, in nodo che includa CpG che ha una certa probabilità di transizioni con una probabilità di osservare un sito attivo più la probabilità di osservare un ATC poi la probabilità di giunsione esone-introne. In teoria possiamo costruire un modello statistico del gene che permette di osservare tutto quello che abbiamo visto fino ad ora. Questo è uno dei metodi che hanno permesso di studiare il genoma umano. Tutte queste probabilità sono ricavate contando quello che succede nei geni che già conosciamo. Il modello di Markof può essere costruito per tutto quello che si vuole. Purché si abbia un sufficiente numero di campioni nel Traning Set.

1.4 Sensibilità, Specificità e curva ROC

Abbiamo visto che vari geni possono essere rimessi insieme attraverso un metodo statistico come i modelli nascosti di Markof. Con questi modelli vorremo essere in grado di predire la posizione di un gene sul genoma. Tutte le volte che sviluppiamo un metodo di inferenza, che dia come output una categorizzazione, abbiamo bisono di un modo per valutare quanto bene funziona il nostro metodo.

Per fare ciò abbiamo bisogno di un set di dati (traning set) da cui stimiamo i parametri del modello, e un test set, insieme di dati di cui conosciamo la risposta, che non sono stati utilizzati per fittare i parametri del modello, su cui testare se il metodo funziona. Supponiamo di voler categorizzare la presenza di una sequenza prima del gene. Vogliamo vedere se il metodo ideato funziona correttamente. I due set devono essere il più diversi possibile; una volta sviluppato il metodo dobbiamo costruire una matrice di confusione. Conosciamo nel test set i casi positivi e quelli negativi, se il metodo identifica correttamente un gene otteniamo un True Positive (TP), se il metodo riconosce un gene non presente otteniamo un False positive (FP), poi abbiamo i TN (True Negative - il metodo riconosce correttamente l'assenza del gene) e i FN (False Negative - il metodo non riconosce la presenza di un gene).

Se sviluppassimo un metodo che identifica tutti come un gene avremo il 100 % di veri positivi ma tantissimi falsi positivi. Quindi dobbiamo mettere insieme questi numeri per ottenere una stima di quanto buono è un metodo[2].

Il parametro di Sensibilità ci dice quanto è sensibile il metodo (in che percentuale indovina i veri positivi), la specificità ci dice quanto è specifico (in che percentuale indovina l'assenza di un gene).

$$S_b = \frac{TP}{TP + FN} \qquad S_p = \frac{TN}{TN + FP}$$

I parametri più usati sono Sensibilità – Specificità e accuratezza. Si usa spsso il coefficiente di correlazione di Matthews che più è alto meglio è (MCC):

$$MCC = \frac{(TP \cdot TN)(FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}}$$

Vogliamo un metodo che sia molto sensibile e molto specifico. Durane la seconda guerra mondiale c'erano gli operatori Radar, dove c'è questa cosa bluastra, con il puntino luminoso. Il puntino luminoso può essere un aereo che sta per attaccare, ma anche un errore. Per poter valutare la qualità o l'accuratezza degli operatori radar si sono inventati la curva ROC (ricevitore operatore), che mette in relazione la specificità e sensitività. Più questa curva è

[2] Un metodo perfetto ha il 100 % di TP e lo 0 % di FP.

alta meglio è. Si usa l'area di questa curva per dare una stima di quanto è buono il metodo. In realtà quello che interessa davvero è cosa accade nella parte basso a sinistra (al venti percento) della curva (Figura 1.4).

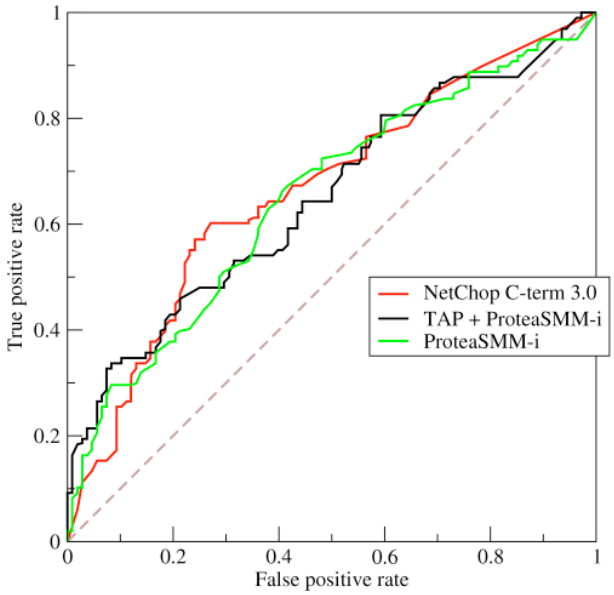


Figura 1.4: Esempio di curva ROC

Capitolo 2

Allineamento di sequenze

Il pesce Fuku ha un DNA particolarmente adatto a questo tipo di studi poiché il suo genoma ha introni molto corti, per cui è più semplice individuare in lui i geni. Abbiamo bisogno però di sapere se due sequenze sono o meno omologhe (quale regione corrisponde evoluzionisticamente allo stesso gene). Confrontare sequenze è una cosa che viene utilizzata per i geni. Sono rarissimi gli articoli che hanno a che vedere con oggetti biologici che non trattano di allineamento. Bisogna vedere se due sequenze sono simili, quanto sono simili, e in cosa corrispondono.

Ad esempio, se prendo due sequenze proteiche e le voglio allineare, cerco di mettere gli aminoacidi della proteina e gli aminoacidi dell'altra in modo da ottenere una sovrapposizione. Se questo è possibile sto implicando che evolutivamente le due proteine provengono l'una dall'altra. Come si fa a decidere se due proteine provengono da un genitore comune? Due proteine hanno una certa probabilità di essere omologhe se la loro probabilità di essere simile è maggiore a quella che io mi aspetti per caso.

Cosa vuol dire che due proteine sono evolutivemanete correlate? I geni dei due individui che producono le proteine vengono da un gene ancestrale di una popolazione che poi ha speciato (si è differenziato). Quanto è probabile che due sequenze sia derivate da un organismo? Le proteine sono formate da 20 aminoacidi, voglio sapere se due sequenze che osservo sono più simili tra loro di quanto mi aspetterei per caso.

2.1 Distanza

Il primo passo è quello di definire una "distanza". Il modo più banale è quello di dire quanti aminoacidi nella sequenza sono diversi. Prima però dobbiamo trovare l'allineamento, ossia trovare la corrispondenza, aminoacido per aminoacido, tra le due proteine.

Supponiamo di averlo trovato, qual è la probabilià di osservare la corrispondenza nelle due proteine degli aminoacidi x e y dato il modello M? Qual è la probabilità di osservarla per caso? Il prodotto della frequenza di x per la frequenza di y, nell'ipotesi di indipendenza. Se ho un modello diverso mi aspetto che la probabilità di osservare questo allineamento sia il prodotto della probabilità di osservare ciascuna coppia nell'allineamento. La distanza tra due sequenze si può definire come il rapporto tra quello che osservo e quello che mi aspetto per caso. Ottengo il prodotto della probabilità di osservare xy e diviso le frequenze di x e di y, siccome siamo interessati ad avere grandezze additive facciamo il logaritmo.

$$d = \sum_{n=1}^{N} \ln\left(\frac{p(x_n, y_n)}{p(x_n) \cdot p(y_n)}\right)$$

Per contare quanti aminoacidi sono ugauli si usano le matrici di sostituzione, si mettono righe per colonne gli aminoacidi, e sulla diagonale ci sono aminoacidi uguali, quindi il punteggio 1 se gli aminoacidi sono identici 0 se sono diversi.

Questa non è una grande idea, si possono prendere i 20 aminoacidi e si può fare una caratterizzazione, se si atribuissi dei punteggi diversi a sostituzioni diversi di aminoacidi un po più complessa. La matrice di sostutizioni sono due numeri che riflettono i logaritmi della probabilità che i numeri di sostituirsi l'uno con l'altra. Se questa fosse la matrice identità sarebbe il metodo banale di prima. Questa matrice mi da un valore numerico della distanza evolutiva tra ciascun aminoacido.

Tutto questo è nato dal laoro di una signora, Margaret Dayhoff ha preso un certo numero di coppie di sequenze, ha preso coppie di proteine che avevano in media una differenza ogni 100 aminoacidi. L'allineamento in questi casi era ovvio. Adesso lei si calcolò la frequenza della sostituzione dell'aminoacido a e b, e ha attribuato un punteggio alla coppia:

$$s(a, b) = \ln\left(\frac{p_{ab}}{q_a q_b}\right)$$

Lei chiamò questa cosa PAM. Il punteggio di osservare la variazione di ciascun aminoacido e altri aminoacidi. Questa matrice di sostituzioni va bene per proteine che si trovano a distanza di 1 aminoacido ogni 100, per avere matrici tra proteine a distanze mattiori occorre moltiplicare tra di loro queste matrici. PAM250 per ciascuna coppia di aminoacidi è correlato alla probabilità di osservare quella variazione tra due proteine che sono a distanza di 250 mutazioni accettate ogni 100 aminoacidi. Ci sono delle approssimazioni, la prima è che stiamo mettendo insieme tutte le proteine, l'altra cosa è che osservare la mutazione non ha nessuna influenza con quello che c'è prima o dopo (e questa è molto forte). L'altra serie di matrici che vengono usate sono quelli di BLOSUM, che sono ricavati da allineamenti di famiglie di proteine. Queste matrici sono ricavate da regioni conservate di famiglie, hanno anche loro un numerino diverso delle matrici pam, Blosum sono all'inverso, più è basso il numero più sono adatte ad allineare le mutazioni in muteini che hanno il 50 % di identità (BLOSUM50).

Salvo che se ho un allineamento come si calcola la distanza? Sommo i valori delle mtrici per le coppie di aminoacidi degli allineamenti, durante le evoluzioni ci sono anche molto insersioni o delezioni.

Insersioni e delezioni sono rare durante le evoluzioni, ci aspettiamo che siano meno probabili delle sostituzioni. Se voglio stimare la similarità tra due sequenze. Ci sono vari modi in cui posso decidere di penalizzare i punteggi, ad esempio ogni volta che c'è un insersione sottrago un certo valore. Oppure si può contare il numero di aminoacidi inseriti e togliere un valore per quel numero. Oppure sottraggo un certo valore quando ho un inserssione, e tengo conto della lunghezza dell'insersione di ogni aminoacido un po' meno. $y(g) = -g + (d-1)e$ (penalizzazione affine).

Le proteine sono strutture molto compatte, le posizioni in cui si possono inserire aminoacidi. Questi valori di g e d vengono settati euristicamente, facendo tante prove e vedendo come viene meglio l'allineamento. Qual è l'allineameno corretto, come facciamo ad avere un test-set di cui sappiamo la risposta?

2.2 Allineamento

Le strutture tridimensionale delle proteine sono molgo meglio conservate degli allineamento, Queste coppie di proteine di cui conosco la struttura tridimensionale, perché so in quel caso quali. Abbiamo le sequenze delle proteine, una penalizzazione che mi permette di includere in questo calcolo. Ci serve un algoritmo che date due sequenze ci trovi qual è l'allineamento ottimale, che meglio rappresenta la relazione evolutiva tra due proteine. Qual è la sequenza che più probabilmente corrisponde allo stesso aminoacide di un progenitore comune?

Qual è il modo di mettere in corrispondenza queste sequenze in modo da massimizzare la probabilità? Vogliamo mettere il più possibile in corrispondenza coppie di aminoacidi che hanno alto valore nelle matrici. Questo problema si risolve in un algoritmo che si chiama programmazione dinamica, ricorsiva.

Si vogliono allineare due sequenze, mettere gli aminoacidi di una con gli aminoacidi dell'altra, facendo in modo che gli aminoacidi corrispondenti si combinino.

L'allineamkento corrisponde ad una linea spezzata in una matrice. Tutte le volte che c'è un segmento orizzontale è un insersione nella prima sequenza, se la linea è verticale3 ho un insersione nella seconda sequenza. La linea ideale è il percorso in questa matrice che ha punteggio più alto?

Riempiamo le celle di questa matrice, con numeri presi dalle celle di sostituzione , quindi il miglior percorso è quello che passa per le celle più alte.

L'algoritmo di programmazione dinamica può essere utilizzato in qualunque caso abbiamo stringhe e voliamo trovare la massima similarità, purché abbiamo qualche punteggio da attribuire alla coppia. Le righe e le colonne rappresentano le insersione e delezioni. Vado in ogni cella, e calcolo il valore massimo passando da quella cella, o venendo da sopra (gap), o da sinsitra (gap), o in diagonale (sostituzione). Il valore della matrice nella cella ij è il massimo (o minimo) tra quei valori. A questo punto riempiamo la matrice in questo modo. La similarità di queste due sequenze è la posizione finale della matrice, siccome io mi ero salvato i percorsi posso ricostruire all'indietro i percorsi.

In ogni cella della matrice metto il massimo punteggio che posso ottenere venendo da una delle celle vicine. Il punteggio dell'ultima cella è il punteggio finale del miglior allineamento.

C'è una piccola modifica dell'algoritmo, Smith e Waterma, che permette di trovare un allineamento miglire locale. Miettiamo a zero tutti i valori negativi, e vedieamo tutti i valori positivi con il valore più alto e si ferma appena incontra uno zero in questo modo trovo il miglior allineamento globale.

Con gli acidi nucleici è esattamente la stessa cosa. In generale per loro o si usa la matrice unitaria. Le basi non hanno molte caratteristiche che possiamo usare per derivare le matrici di PAM.

2.3 Famiglie di proteine omologhe

A questo punto abbiamo le due sequenze allineate. Ho una sequenza, come posso torvare se esistono sequenze omologhe? Questo mi permette di inferire la catena evolutiva oltre che aiutarmi a capire sia la funzione di alcune proteine (presumibilmente proteine omologhe avranno funzioni simili) che la struttura tridimenzionale. La sequenza da cui partiamo si chiama sequenza *query*.

Allineo la *query* con tutte le sequenze delle banche dati, e calcolo il pnteggio di similarità.

Bisogna controllare se all'interno delle decine di milioni di sequenze che conosco ce ne sono alcune tanto simili alla mia sequenza query da permettermi di asserire con una certa probabilità che sono derivate dallo stesso progenitore comune.

Si prende la sequenza query, si allinea con ciascuna sequenza nota e si calcolano i punteggi. Si fanno alcune approssimazione per evitare di fare troppi allineamenti: ad esempio se non hanno neanche una coppia di aminoacidi identici non si allineano. Lo stesso discorso si applica alle sequenze geniche.

Avremo per ogni sequenza della banca dati un valore del punteggio di similarità. Per capire se questo punteggio è alto o basso occorre fare un po' di statistica. Per valutarne la significatività si confronta con il punteggio medio dell'allineamento di due sequenze completamente casuali. La statistica di quello che ci si aspetta è risolta analiticamente nel caso di allineamenti locali senza gap. La distribuzione che mi aspetto in questo caso è la distribuzione dei *valori estremi*.

2.3.1 Distribuzione dei valori estremi

Se si va in una scuola e in ogni classe si prende il valore dello studente più alto si ottiene la distribuzione dei valori estremi.

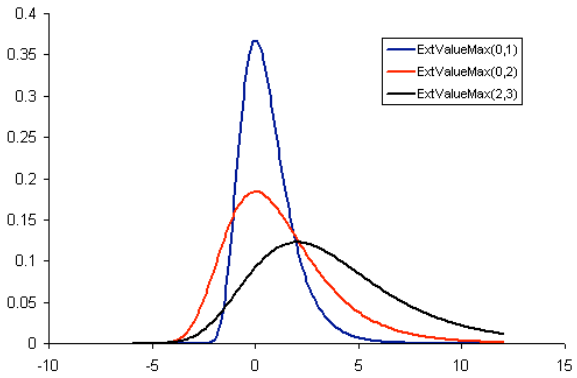


Figura 2.1: Grafico della distribuzione dei valori estremi.

Questa distribuzione è asimmetrica verso i valori più alti (Figura 2.1). Si può dimostrare che se i segmenti senza gap sono simili e di lunghezza sufficientemente grande, il numero di volte con cui mi aspetto di osservare coppie con punteggio di similarità S è pari a:

$$E = Kmne^{-\lambda S}$$

Dove m ed n sono le lunghezze dei segmenti, K e λ sono fattori di scala. Questo numero si può usare anche quando parliamo di allineamenti globali. Se si vogliono confrontare valori ottenuti con metodi diversi si può cercare di elimiare questi λ e K. Si può usare come punteggio di similarità un altra versione:

$$S' = \frac{\lambda S - \ln(K)}{\ln 2}$$

Con questa definizione il valore atteso del numero di casi in cui osservo il punteggio S' che non dipende più dal metodo usato:

$$E = mn2^{-}S$$

Questo serve se si usano metodi differenti per stimare gli allineamenti. In questo modo si può immediatamente vedere qual è il numero di votle in cui aspettarci questo punteggio per caso.

Il programma ricerca in banca dati la similarità ad una sequenza data, e restituisce ordinatamente le sequenze della banca dati con il relativo punteggio rispetto alla sequenza. Se voglio sapere se il punteggio assegnato è alto o basso, il programma rimescolando la sequenza ripete la ricerca in banca dati.

I punteggi che mi verranno in questa ricerca saranno punteggi per proteine non omologhe. Questo si fa tante volte per generare una statistica, che punteggi mi aspetto casualmente. Queste banche dati ci danno un istogramma quante volte si ottiene un dato punteggio nella ricerca in banca dati casuale (che non ha significato biologico). Quello che viene fuori è una distribuzione di valori estremi.

Questa è la distribuzione dei valori di punteggio che ci si aspetta per caso. Si fa il calcolo euristico dello zscore:

$$zscore(s) = \frac{S - \mu}{\sigma}$$

Se il punteggio è zero è molto probabilie che questo punteggio non dia un omologia, se avessi un punteggio che sta molto fuori dalla distribuzione casuale ho una certa probabilità che il mio punteggio sia dovuto non al caso ma a qualcosa diverso (come l'evoluzione).

Lo zscore è banalmente la distanza tra il punteggio e la media della distribuzione casuale misurato in termini della deviazione standard. Questo mi permette di calcolare quante volte mi aspetto di osservare questo zscore in una ricerca in banca dati.

In genere si fa un taglio dello zscore molto largo, e ci sono una serie di sequenze che definiamo omologhe ed una serie di sequenze che non consideriamo omologhe. Avrò o veri positivi che sono le sequenze omologhe che ho tagliato, e posso fare la solita curva Roc tra i programmi per cercare. BLAST è il programma che ha la migliore curva ROC e velocità.

BLAST è veloce perché imbroglia: non ripete le ricerche casuali tutte volte, ma usa una statistica preconfezionata abbastanza ragionevole. Oltre a dare la lista della sequenza con il punteggio il programma fornisce anche l'allineamento ottimale. Si possono avere dei metodi più sensibili o più specifici.

Che valore di E voglio utilizzare come taglio? Dipende dalle esigenze. Nelle banche dati gli errori si propagano. Se ci sbagliamo a categorizzare una proteina, e asseriamo che una proteina della rosa sia la distrofina perché è molto simile alla distrifina umana, se ci sbagliamo il prossimo tizio che trova una proteina simile alla nostra, otterrà che anche questa è una distrofina, propagando l'errore.

Se si fanno gli assegnamenti automatici delle funzioni si deve usare valori di E molto selettivi, se invece si ha intensione di andare a analizzare manualmente tutti i match possono prender e E minore.

2.4 Allineamenti Multipli

Fino ad ora abbiamo sempre allineato coppie di sequenze, si può anche pensare di fare un allineamento multiplo. Cioè mettere in corrispondenza gli aminoacidi di sequenze omologhe, allineandole assieme. Perché è utile fare gli allineamenti multipli? Perché siamo in grado di capire meglio la funzione delle proteine: gli aminoacidi conservati in tutte le proteine omologhe sono i più importanti, perché quando sono mutati hanno distrutto la funzione della proteina. Se due sequenze sono talmente vicine che presumibilmente qualcuno di quegli aminoacidi non è mai cambiato, in compenso quando allineo sequenze molto simili l'allineamento è banale.

Se vogliamo allineare due sequenze differenti possiamo però allineare sempre coppie di sequenze abbastanza vicine e osservarte alla fine l'allineamento migliore a sequenze distanti. Dall'avere allineato tutta la famiglia ottengo informazioni aggiuntive in quanto in una coppia è possibile che alcuni aminoacidi siano uguali per caso, se però lo sono in un intera famiglia allora il dato è più significativo.

Potrei usare lo stesso algoritmo per allineare un intera famiglia, usando la programmazione dinamica, tuttavia la complessità del problema è troppo alta. Esistono algoritmi euristici per arrivare alla soluzione.

Algoritmo iterativo

Allineo due sequenze, poi allineo una terza con l'allineamento delle prime due, e così via.

Per allineare una sequenza con un allineamento di due sequenze, al posto dei punteggi ottenuti dalle matrici di sostituzione si usa la media dei punteggi sulle sequenze allineate in quella posizione.

L'unica cosa che occorre decidere è da dove si comincia? Conviene allineare prima le sequenze più simili perché sono più facili da allineare.

Come si trovano le sequenze più simili, questo serve anche per operazione di clusterizzazione. Un modo per clusterizzare è proprio questo.

La prima cosa da fare è allineo ciascuna sequenza con ciascun'altra sequenza (si mette sottoforma di distanza). Si prende la coppia più simile e la allineiamo.

Si costruisce una nuova matrice di distanze, dove AB è come se fossero un unico elemento (ho allineato A e B per primo). La distanza con le altre sequenze si ottiene come media tra le distanze di A con C e B con C (e questa è la distanza di AB con C). Si ripete esattamente lo stesso procedimento, ad esempio se la distanza minore è con D.

Si possono rappresentare le distanze come un albero, e questo è visibile come un albero evolutivi. Per fare un vero albero evolutivo la cosa è un po' più complessa, tuttavia è utile per fare i raggruppamenti tra gli oggetti.

Center Star

Esiste un altro metodo, detto di Center Star. È banale, abbiamo le sequenze, si calcola la distanza fra tutte le coppie di sequenze, e ottengo una matrice simmetrica (con la similarità al posto della distanza). Si cerca la sequenza che è la più simile in media a tutte le altre. La sequenza per cui la somma delle distanze dalle altre è minima. Se scelgo la sequenza più centrale, allineo tutte le sequenze con S1. SI prende il primo allineamento e si aggiunge la sequenza successiva allineandola alle precedenti due già allineate, ma tenendo conto del suo allineamento originale con S1.

2.4.1 Profili

I profili sono matrici sitospecifiche di una famiglia di proteine. È la tabella di probabilità di osservare un certo aminoacido in una certa posizione. Si fa esattamente la stessa cosa che si faceva per le matrici di sitospecifico. Si conta quante volte capita un certo aminoacido in una certa posizione in un allineamento multiplo, e si approssima la frequenza con la probabilità, aggiungendo gli pseudoconteggi (poi si fanno i logaritmi).

Gli pseudo conteggi possono essere scelti sia da una distribuzione uniforme, che da una distribuzione non uniforme, facendoli ad esempio proporzionali alla composizione in aminoacidi di un'organismo.

Dato un allineamento avremo nella prima riga la posizione, nelle colonne i vari aminoacidi. E nelle celle metto le volte che ho osservato un certo aminoacido.

Una volta individuata una famiglia di proteine omologhe alla mia con BLAST, si fa l'allineamento multiplo, si ricava un profilo, e questo profilo può essere usato per cercare altre proteine omologhe all'interno del database, in quanto il profilo è un sistema molto più potente del singolo allineamento. Questa cosa si può fare e la fa il programma direttamento, questo la fa il programma PSI-BLAST.

Qui la statistica fa davvero acqua da tutte le parte. Anche qui non c'è una teoria statistica che ci aiuta molto. Fa un loop costruisce un allineamento multipli (MSA) genera un profilo cerca nuove sequenze, e itera questo processo fin quando non arriva a convergenza.

Come si fa quando si ricercano segnali con le matrici sito-specifico, il passo successivo è quello di passsare dai profili ai modelli nascosti di Markov.

Supponiamo di avere una famiglia di proteine allineata. Questo allineamento non solo può essere descritto con un profilo, modello nascosto di Markof di ordine 1. In questo modo posso creare un modello statistico della proteina data.

Cosa ci facciamo con le proteine omologhe? Si vorrebbe poter attribuire una funzione ad una proteina di cui non so nulla. Questo non è banalissimo.

Da un gene ancestrale a causa di una mutazione si sono originate due diverse proteine in due specie diverse (ortologhe), allora queste proteine molto probabilmente hanno conservato la stessa funzione. Tuttavia se in un organimso abbiamo geni che si diversificano a seguito di una duplicazione genica (paralogi) le due proteine continuano a coesistere nel medesimo organismo, per cui la nuova proteina generata (omologa alla prima) potrebbe avere una funzione differente.

È importante avere l'intero genoma, per capire se ci sono state delle duplicazioni. L'assegnazione di funzione per omlogia è molto pericolosa.

2.5 Metodi di classificazione e clustering

Abbiamo un insieme di oggetti e vogliamo raggrupparli in modo che gli oggetti di un gruppo siano più simili tra di loro di quanto non lo siano con gli altri.

Queste sono tecniche sono usate anche nel riconoscimento di immagini, metodo di apprendimento automatico, analisi delle immagini e recupero di informazioni e bioinformatica.

Abbiamo visto come si costruisce un albero con le matrici di distanza. In realtà possiamo ottenere dei cluster a partire dagli alberiFigura 2.2: tagliando l'albero ad altezze diverse si possono ottenere un diverso numero di cluster. Un albero può essere usato per raccogliere dati unsupervised, o per capire come si è evoluta una certa regione genica.

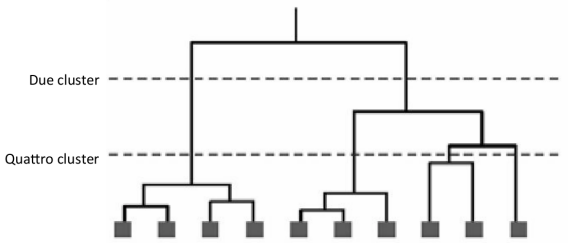


Figura 2.2: Esempio di clusterizzazione a partire dagli alberi.

Quello che facciamo di fatto è allineare le sequenze cluster per cluster. Due proteine diverse sono soggette a pressione evolutiva diversa, proteine che variano molto poco nel corso dell'evoluzione saranno proteine fondamentali per la sopravvivenza dell'organismo. La pressione evolutiva dipende dalla sua funzione, se due proteine interagiscono (formano un complesso) ci aspettiamo che evolvono insieme. Per vedere se effettivamente due proteine interagiscono si prendono gli allineamenti multipli delle due proteine (prendendo le specifiche proteine in specie differenti) e si fanno gli alberi evolutivi tra le due specie. Se gli alberi sono simili vuol dire che le proteine hanno mutato assieme e che quindi presumibilmente interagiscono tra di loro.

Vediamo qualche caratteristica dell'albero. Gli alberi possono essere di due tipi, con o senza una radice. Gli alberi hanno una topologia, ossia possiamo definire una distanza tra i nodi (o le foglie). Normalmente i nodi si etichettano a partire dalle foglie fino a risalire.

Se ci sono n nodi ci saranno 2n-2 connessioni, per un albero con n nodi ci sono n-1 nodi nella parte superiore e n fogli e quindi 2n-2 connessioni. Se abbiamo un albero senza rdice possiamo mettere la radice su ciascun delle connessioni. Siccome ci sono 2n-3 connessioni, per ogni albero senza radice possiamo ottenere 2n-3 alberi con radice.

Il metodo più semplice per costruire un albero è quello che abbiamo visto l'altra volta: il metodo UPGMA, che si ottiene mettendo i più vicini a coppie.

Se volessimo usarlo per valutare per evoluzioni non è particolarmente adatto perché si può immaginare che la velocità con cui si sono evoluti i geni sono diversi. L'UPGMA si può usare solo se la distanza è tale per cui si è una situazione in cui la distanza corrisponda al tempo. L'altra assunsione che facciamo con l'UPGMA è che le distanze siano additive.

Se assumiamo comunque che le distanze siano additive ma nessuna assunzione sull'accumulo di variazioni si può usare il metodo *Neightbour joining*. Ho delle distanze date, voglio costruire un albero che abbia dei nodi in più in modo tale che la distanza sia allineata. Esistono degli algoritmi per individuare in che posizione dobbiamo aggiungere i nodi.

Un altro modo per costruire alberi è il metodo della *massima parsimonia* (spiegare i dati con il minimo numero di approssimazioni possibili). Dobbiamo trovare una struttura che faccia il minimo numero di sostituzioni possibili.

Si può calcolare il numero di sostituzioni che si deve fare per passare da un nodo ad un nodo figlio. Il numero di sostituzioni si chiama costo dell'albero. Dovremo costruire tutti i possibili alberi, e cercare quello di costo minimo. Questo approccio detto di Forza Bruta è inutilizzabile. Ci sono dei trucchi che in alcuni casi garantiscono di trovare l'approsisimazione ottimale, in altri casi no.

Potremo costruire un albero a caso, e cominciare ad invertire due nodi dell'albero, se diminuisce vado avanti, altrimenti torno indietro. Questo metodo mi esclude la possibilità di non trovare minimi locali. Prendiamo tre sequenze a caso, cominciamo a costruire un albero, e vediamo dove ci costa meno aggiungerla una quarta. Anche questa è una specie di minimizzazione, e anche in questo caso vado su un minimo locale.

Il metodo più usato è il Branch and bound. Questo metodo funziona se il costo è additivo (anche l'energia è additiva). Si iniziano a costruire i possibili alberi. Se ad un certo punto il costo di un albero mi supera quello di un altro albero completo, allora questo albero che sto costruendo sicuramente non è ideale, e posso bloccarmi.

2.5.1 Bootstrap

Supponiamo di aver costruito l'albero, vorremo sapere quanto è affidabile l'albero. È l'unico albero che possiamo costruire? Facciamo il bootstrap. Abbiamo l'allineamento multiplo con una distanza fra le varie sequenze. Tra queste distanze abbiamo ottenuto un albero.

Adesso prendiamo l'allineamento multiplo e selezioniamo a caso 100 delle colonne di allineamento multiplo, con possibilità di selezionare due volte anche la stessa sequenza. Ricostruiamo l'albero da questo insieme, e ripetiamo questa simulazione tante volte

Quante volte le proteine sono raggruppate insieme? Per ogni ramo posso mettere un numero che mi dice quanto spesso gli alberi riproducono quella caratteristica. Più volte le proteine sono raggruppate nello stesso modo, più vuol dire che quel cluster è significativo.

2.5.2 Uso della clusterizzazione per diagnosticare patologie

Supponiamo di sequenziare tutti i DNA della cellula tumorale e quelli nella cellula sana. Questo ci può servire per varie applicazioni. Gli RNA che fanno la proteina hanno una codina di adenina che possono essere utilizzate per estrarli dalla cellula e poi possono essere analizzati: Si prende il DNA di una certa popolazione, lo si amplifica e si fa fluorescere nel verde. Poi prendo il dna dall'altro tipo cellulare e mando la fluorescienza nel rosso. In un vetrino dove pongo le seuqenze complementari a tutti i geni umani che conosco. Prendo il campione verde rosso e li mischio e li deposito su questo vetrino. Succederà che le molecole si legheranno al gene complementare. E vedo giallo se si lega a entrambi, verde se si lega solo quella sana, rossa solo quella tumorale nero se non si legano nessuno.

A questo punto poiché per ogni gene sappiamo i colori, possiamo clusterizzare raggruppando insieme quelli che hanno comportamenti simili (la distanza sarà la differenza di flourescenza). Catalogare i geni in funzione del loro lvello di espressione.

Questo è un modo, ce ne sono altri. I metodi di classificazione o cluster possono essere supervised o altro.

Trovami i geni che distinguono i pazienti con metatstasi e i pazienti senza metastasi, oppure dico questi sono i miei dati clusterizzali.

Metodi supervised sono metodi in cui isnegno ad un algoritmo a riconoscere una caratteristica. Tra i metodi per insegnarlo ci sono delle Reti neurali, Alberi decisionale, Support vector machines.

2.6 Reti Neurali

Una rete neurale è una rete di neuroni che imita quello che avviene in un cervello: Ogni neurone riceve degli imput dai neuroni sottostanti, e invia un output ai neuroni del layer successivo. Se l'integrale dell'inpulso (pesato) è superiore ad una certa soglia manda un output. Ad esempio si può definire una funzione di trasmissione:

$$y = \sum_i w_i x_i - \beta$$

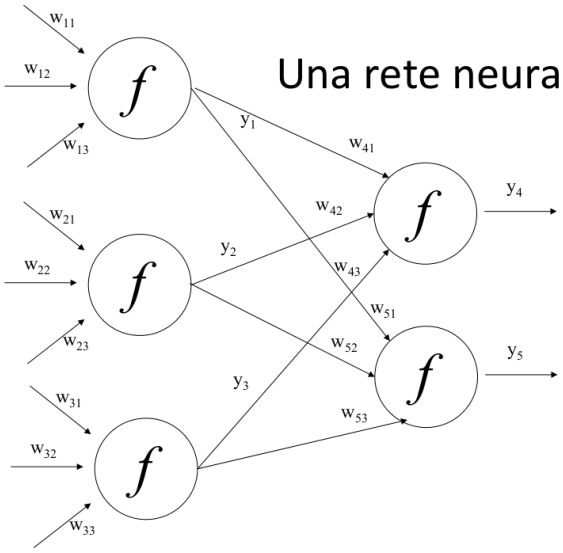


Figura 2.3: Rete neurale

Possiamo costruire una rete neurale unendo insieme questi neuroni, se il valore che ottiene la y è superiore ad una certa soglia invierà un output al neurone che lo segue. La rete per funzionare deve essere addestrata.

Voglio sapere se una proteina è più probabilmente extracellulare o intracellulare. Prendo un insieme di proteine di cui conosco le caratteristiche e che so se sono o meno extracellulari. Costurisco una rete che ha un algoritmo per modificare i pesi dei singoli neuroni in modo che dato un input, mi si modifichino i pesi per avere l'output desiderato. Alla fine dell'addestramento la rete avrà imparato quale combinazione di pesi è in grado di distinguere meglio i due casi.

Dopo aver fatto tutto questo non non sappiamo che cosa la rete ha imparato realmente. Lui userà i parametri che io gli do per distinguere i set di dati. Che cosa ha imparato il mio modellino? Per sapere quanto bene ha imparato occorre fargli un esame, cioè devo avere un altro set di dati per i quali conosco la risposta e lui non ha mai visto. Devo fare in modo che il test set sia simile al training set solo per la caratteristica che mi interssa. Posso anhe prendere il training set, ne tolgo un decimo e faccio il training sui 9/10 e vedo come funziona, rimescolando a caso ogni volta. Un altra cosa il testing set si butta dopo essere utilizzato, si usa solo ed unicamente una volta. Il test set deve essere buttato una volta usato.

Le funzioni di trasferimento possono anche non essere lineari nel peso. Lo strato di neuroni a cui si danno i dati è input-layer, abbiamo l'output-layer, e poi abbiamo gli stati di neuroni finali su cui non interagiamo, l'hidden layer è sopra. In pratica, quello che fa una rete neurale, è quello di trovare una retta che separi i dati positivi dai dati negativi.

Si può studiare la funzione di errore in funzione della variazione dei vari parametri, o variando il peso di un parametro se ha un efftto sul risultato.

2.6.1 Uso delle reti neurali per predire la struttura delle proteine

Se si guarda la struttura di una proteina ci sono delle regioni (strutture secondarie) che sono strutture locali ripetitive, che sono regioni delle proteine che hanno caratteristiche particolari. La rete neurale può essere usata per inferire se un certo aminoacido si trova in una regione α elica o in un filamento β? Si prendono proteine di cui si conosce la struttura, facciamo allineamenti multiple e ricaviamo il profilo di questi allineamenti omologhi. Sappiamo qual è una regione che corrisponde ad una α elica. Prendiamo una finestra e chiediamo alla rete di predire la struttura dell'aminoacido centrale, poi si fa il training della rete neurale.

Abbiamo 13 input, Ad ogni input corrisponde un array di venti numeri. Ciascuna posizione dell'array rappresenta il numero di volte in cui quell'aminoacido compare nel profilo.

L'output che vogliamo è se l'aminoacido centrale si trova in una alpha-elica o in un filamento beta. Poi spostiamo la finestra in modo da inferire la posizione di tutti gli aminoacidi centrali.

Ora su una proteina di cui non so nulla devo prima trovare tutte le proteine omologa alla mia, costruire un profilo, e la rete mi farà una predizione degli elementi che sono delle regioni α e β della proteina. Questo metodo ha una accurattezza superiore al 90 %.

Se usassimo dei numeri per codificare gli aminoacidi la rete potrebbe imparare qualcosa di sbagliato, ad esempio che l'aminoacido codificato con 9 è motlo più "vicino" dell'aminoacido codificato con 10 pituttosto che con quello codificato con 18. Poiché questa informazione non ha alcun senso biologico, può essere dannoso passarla alla rete (ecco perché si usano i vettori a 20 dimensioni)

Cosa vuol dire utilizzare una rete neurale? Una rete in pratica trova qual è la retta che mi possa separare i dati, dopo che sono stati separati sotto un opportuna interazione. Tuttavia se i dati non sono tanti, o lo spazio in cui si distribuiscono è molto grande, esistono tantissime rette possibili. Per trovare quella migliore occorre usare la *support vector machine*

Un modo semplice per pensare di separare i dati può essere quello di trovare il poligono convsso che racchiude tutti i dati di un tipo da quelli di un altro. Trovo la linea più breve che congiunge i due poligoni, e la retta che scelgo per separare i dati è l'asse di questo poligono. In questo modo si ha la miglor divisione dei dati. Supponaimo che i dati da separare siano difficilmente separabile da una sola retta. Però se si prova a fare una trasformazione di coordinate si può trovare una maggiore separabilità dei dati. Questa trasformazione può anche essere non lineare, e prende il nome di Kernel.

2.7 Random Forest

Il metodo che piace a tutti è la *Random Forest*. Il nome viene dal fatto che si utilizza una foresta di alberi decisionali, e si insegna a questi alberi a separare i dati. Quando gli alberi sono addestrati si passa un dato, e la foresta vota. La maggioranza vince.

2.7.1 Alberi decisionali

Un albero decisionale è un albero su cui in ogni nodo c'è una decisione e le foglie sono gli esiti della decisione (valutate in base agli input). Questi alberi possono essere fatti in vario modo.

Dato un albero posso vedere quanto bene l'albero ha funzionato. Dobbiamo avere un metodo per capire se un albero, sia nella domanda che nella decisione, cataloga correttamente.

L'indice di *Gini* ci da questa informazione. Si prendono i dati e si dividiono in n classi, e ci si chiede quanto sono pure le classi rispetto alla suddivisione per un certo attributo:

$$I_G(g) = \sum_{i=1}^{m} f_i(1 - f_i) = 1 - \sum_{i=1}^{m} f_i^2$$

Dove f_i è la frazione di elementi aventi l'attributo che desidero fanno realmente parte di quella classe.

Più è basso l'indice di gini, miliore è la purezza delle classi. Possiamo addestrare gli alberi decisionali in modo che mi minimizzino l'indice di Gini per il mio training set.

Per studiare casi in cui le variabili evolvono in modo continuo devo prima partizionare la variabile. Prendere il valore medio della partizione, e catalogare quanti casi sono nella prima e quanti nella seconda metà della partizione. Calcolo l'indice di gini delle due metà e poi lo medio pesandolo sul numero di casi.

Posso generare diversi alberi decisionali, e si può calcolare l'indice di gini per ciascun albero e decidere qual è l'albero milgiore.

Facciamo una foresta di alberi. Abbiamo i dati e un certo numero di variabilìì: N casi del training set e m variabili da cui dipende la decisione. Seleziono i dati con rimpiazzo (posso ripescare alcuni dati più di una volta) e generiamo gli alberi. Per ogni nodo che compone l'albero scegliamo a caso una delle m variabili, che usiamo come decisione di quel nodo. Avremo tanti alberi che danno una decisione, e sono stati costruire usando dati noti.

A questo punto si va con il dato ignoto, a ciascun albero, e vince la maggioranza dei risultati. Uno dei vantaggi è che non è essensiale un test set. Supponiamo di avere dei classificatore ogniuno dei quali indovina il 70 %. La probabilità di avere la risposta giusta, con 101 classificatori è del 99.9 %.

Il Random forest permette di dire se si ha classificato bene, ci può dire anche qual è la l'albero che sta funzionando meglio all'interno del Ramdom forest (e quindi quali variabili sono più determinanti). Bisogna fare attensione all'overfitting, se uso un piccolo training test non bisogna esagerare con i parametri, altrimenti il sistema imparerà a riconoscere singolarmente le sequenze singolarmente le sequenze del traing set.

2.8 Principal Component Analisys

La PCA è un metodo a che ha due scopi, serve sia per classificare i dati, che per ridurre la dimensione dei dati e caratterizzarli. La PCA è un metodo per dire in che coordinate i dati danno maggiore informazione. Sto trovando un nuovo sistema di riferimento in cui in un asse c'è la maggiore variazione dei dati, sul secondo una minore ecc. Se ho una coppia di valori x e y posso definire la covarianza, il prodotto tra la varianza di x e quella di y.

$$\text{cov}(x, y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

La covarianza mi dice quanto sono collegati i dati. Possiamo definire la matrice di covarianza come una matrice sulla quale ad ogni elemento corrisponde la covarianza fra le due variabili corrispondenti in riga e colonna.

Bisogna calcolare autovalori ed autovettori di questa matrice. L'autovettore che corrisponde all'autovalore più grande è quello che meglio separa i dati.

Se voglio approssiamare i dati. Se il primo autovalore è molto alto e il secondo è molto basso vuol dire che il primo asse ben mi approssima i dati, e posso trascurare il secondo.

La *Principal Component Analysys* (PCA) cerca di ridurre la dimensione dei dati, si selezionano i vari assi su cui proiettare i dati in modo che siano i più importanti possibile nell'esprimere i dati.

Se tutti gli autovalori sono uguali non c'è verso di semplificare la cosa, vice versa se i primi 4 dati riproducono il 95 % degli autovalori, allora possiamo trascurare tutti gli altri. Guardando gli autovettori, dalla combinazione di colonne e righe della matrice iniziali ha dato luogo a quegli autovettori si può sapere quali sono gli elementi più importanti. Un esempio in cui viene lagamente usata la PCA è quello di misurare quali sono le cellule tumorali e quali sono le cellule non tumorali e si avranno tanti parametri, la PCA ci permette di vedere se c'è un modo di separare i dati nel modo più evidente possibile.

Capitolo 3

Struttura delle proteine

3.0.1 Aminoacidi e strutture secondarie

Le proteine sono formate da aminoacidi, con un gruppo carbonio centrale, detto gruppo α, a cui sono legati gruppi variabili. Gli amioacidi esistenti in natura sono tantissimi, quelli codificati dal nostro genoma sono 20 o 21. Gli aminoacidi che utilizziamo sono tutti *levogiri*.

Mettere insieme aminoacidi levogiri e destrogiri è impossibile, quindi il nostro modo è fatto da aminoacidi levogiri. Alcuni aminoacidi sono idrofobici, non in grado di formare legami idrogeni, alcuni sono carichi positivamente altri sono carichi negativamente, e altri sono polari.

C'è un aminoacido particolare che è la *Cisteina*. Questo aminoacido in ambiente ossidante, se i gruppi SH della cisteina sono nell'orientamento giusto possono formare un legame covalente tra gli atomi di zolfo, liberando gli idrogeni. Questo è l'unico legame covalente che si forma tra gli aminoacidi. La carica di un aminoacidi è dipendente dal pH, se il gruppo amminico e quello carboossidrico si legano o meno agli $H+$ e $OH-$, e quindi dalla loro abbondanza in soluzione. Il carbonio di un aminoacido può essere legato covalentemente all'azoto dell'altro aminoacido (unendo un gruppo amminico al gruppo ossidrico) lasciando comunque un gruppo ammiinico e uno ossidrico, in questo modo polimerizzano. Formando l'intera proteina.

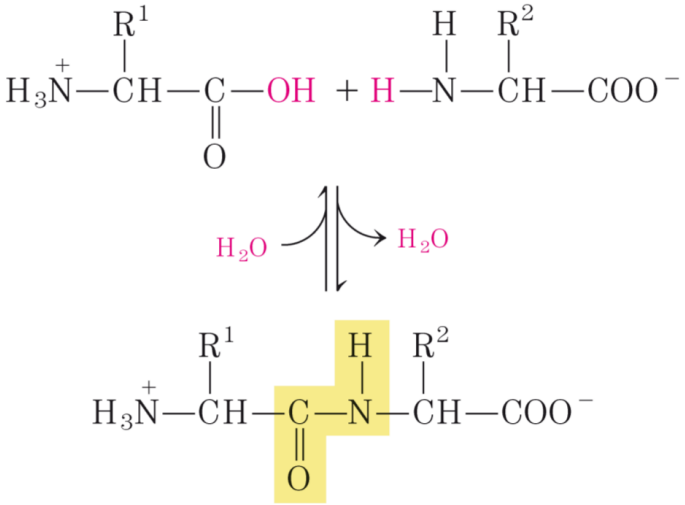
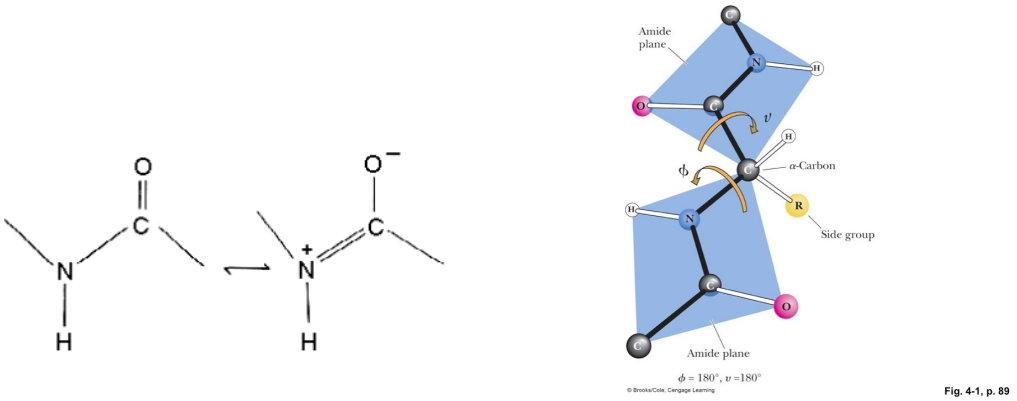


Figura 3.1: Esempio di polimerizzazione.

La specifica sequenza di aminoacidi è determinata dalla corrispondenza di codoni e tripletta del DNA. Ogni proteina ha una sequenza di aminoacidi diversi. Le mutazioni avvengono a livello di Geni, le funzioni avvengono al livello della proteina. In una catena aminoacidica sul carbonio α può legarsi una catena laterale. C'è una parte ripetiva che si chiama della catena principale detta *backbone*. Le catene laterali o residui possono essere polari, neutri, carichi.

Il legame nuovo che si forma tra la catena (Tra azoto e carbonio) ha il carattere di doppio legame parziale (vedi Figura 3.2). È una struttura di risonanza in cui si distribuiscono sia sull'elettrone O che sul legame N. L'angolo diedro intorno a questo legame vale preferenzialmente 180 gradi. H ed O sono dalle parti opposto della catena. Possiamo ruotare la catena laterale intorno al carbonio e abbiamo due angoli diedri che possono variare. Le proteine assumono una forma perché questi angoli ruotano ne variano. (Tra C-N e tra C-C).



(a) Legame parziale doppio (b) Angoli diedri variabili

Figura 3.2: Schema dei legami e angoli tra aminoacidi.

Ruotando intorno a questi angoli non tutte le trasformazioni sono possibili, ci saranno conformazioni molto sfavorevoli. Se ruotiamo varie combinazioni di questi angoli, se questa combinazione è permessa coloro di blu scuro, se non è permessa coloro di giallo. Le regioni in blu e scuro sono quelle combinazioni tali che non ci sono contatti troppo specifici tra gli atomi quando ho scelto quegli specifici angoli. Possiamo ridurre del 10 % la dimensione del raggio di van der Walls e ottenere le regioni in azzurro chiaro (Figura 3.3):

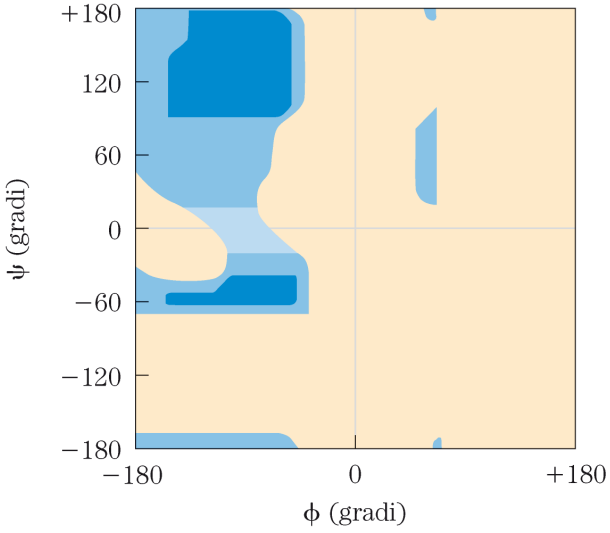


Figura 3.3: Grafico delle combinazioni permesse.

Queste combinazioni di angoli ψ e ϕ sono permesse. La regione gialla è una regione energeticamente molto sfavorita, quella perdita di energia potrebbe essere compensata da altro. Se ci chiediamo che combinazione di angoli ψ e ϕ.

Esistono strutture locali delle proteine che sono particolari posizioni di questo grafico, come le α eliche.

L'alpha elica gli ossigeni e l'idrogeno di un amioacido sono nella distanza giusta per formare un forte legame idrogeno. Questa possibile struttura è stata ipotizzata prima che fosse vista sperimentalmente. Poling capì che il legame NC ha un carattere di doppio legame parziale, la seconda intuizione fu che non necessariamente un elica dovesse avere un numero intero di aminoacidi per giro, con queste ipotesi riuscì a predire la struttura delle α eliche.

La proteina strutturandosi perde un enorme quantità di entropia. Il fatto di formare tanti legami idrogeno significo una perdita di energia, che deve essere compensata in altro modo: le interazioni fra le catene laterali, che se sono più favorevoli nella forma strutturata che nella conformazione libera, mi comportano nun guadagno energetica. Le interazioni possibili sono idrofobiche, di van der walls, elettriche. Una regione si strutturerà ad alpha elica, con le catene laterali che puntano verso l'esterno, se l'interazione delle catene laterali in questa situzione è più favorevole, l'alphna elica si forma spontaneamente. La decisione di formare o non formare l'alpha elica dipenderà dalla specifica sequenza di aminoacidi. La sequenza della proteina localmente è responsabile del fatto che quella regione assuma o non assuma quella regione in apha elica o meno.

Questo è vero per l'elica ed è vero in generale. La proteina raggiunge una struttura unica spontaneamente se e solo se le interazioni sono sufficientemente favorevoli da compensare la perdita di entropia.

Le eliche che osserviamo nelle proteine sono tutte destrogire. L'elica destrogira non è stabile quanto l'elica levogira. L'altra regione del grafico è responsabile di un altro tipo di struttura secondaria che si chiama filamento β.

È una regione abasstanza piatta (angoli vicino a 180). La speranza di ottenere un filamento beta spontaneamente è praticamente nulla, questo perché una catena non ha nessun guadagno di energia. Noi infatti troviamo i foglietti beta, in cui i filamenti si dispongono affianco tra di loro, e si ottengono foglietti che hanno un guadagno energetico rispetto alla stessa sequenza di aminoacidi. I filamenti si possono appaiare in molti modi sia parallelamente che antiparallelamente. Antiparallelamente sono lo stesso filamento che viene ripiegato su se stesso, mentre paralleli sono più filamenti appaiati nella stessa direzione. Se volete riconoscere i foglietti. Nel foglietto antiparallelo i legami idrogeni sono tutti paralleli, in quello antiparallelo sono invece non paralleli.

Le catene laterali possono interagire tra loro il foglietto beta si formerà in base alle energie delle catene laterali.

3.1 Struttura terziaria

La sequenza di aminoacidi è la struttura primaria, può formare delle strutture locali secondarie che poi si assemblano per formare una struttura compatta che si chiama struttura terziaria.

Molto spesso le proteine funzionale assemblano varie catene strutturate e assemblate otteniamo la struttura quaternaria della proteina. Le proteine sono polimeri lineari. C'è un solo caso in cui ci sono legami tra parti diverse della catena. Se due catene si trovano nell'orientamento e alla distanza giusta si forma un ponte e un legame covalente tra due Cistine, detto ponto di solfuro.

Questo è il meccanismo della permanente dei capielli. I capelli hanno delle proteine. Se aggiungiamo un agente riducente rompiamo i ponti di solfuro, mettiamo i capelli nella forma che ci piace, riossidiamo, e le proteine riformano il legame covalente mantenendo la stessa posizione.

Il contenuto di un alpha elica e di foglietti beta può essere valotato usando il *dicroismo circolare*. Proteine formate da alpha elica possono assorbono a diversi spettri di frequenza la luce. Da cui possiamo riconoscere quali componenti ci sono nella proteina.

Esiste una proteina che si chaiama la mioglobina, ed è la prima proteina la cui struttura tridimensionale è stata risolta sperimentalmente.

Questa è formata per la maggioranza da α eliche. Le strutture possono essere molto particolari. C'è un foglietto beta che forma un barile, e queste proteine vengono usate come scambio tra membrane cellulacere. Esistono proteine formate tutte da α eliche, esistono proteine tutte beta, con forme interessantissime, anche eliche di β, elica a quattro pale, sandwich beta, o possono essere miste, α elice + β. Esistono classiche proteine in cui alpha e beta sono impacchettate le une dentro le altre. Molto spesso le proteine sono formate da domini. È come se fossero due nuclei compatti uniti da qualche cosa, le proteine molto grandi generalmente sono formate da domini delle dimensioni di 100-150 aminoacidi, queste proteine hanno maggiore mobilità poiché gli angoli tra i domini sono facilmente modificabili, quindi le proteine che si devono muovere sono spessi generati da domini.

La proteina deve strutturarsi, deve passare dalla sua configurazione strutturale alla configurazione a minima energia. Quello che interessa alla pressione selettiva la proteina ha quella forma per il tempo che serve, ricostruiamo nuova emoglobina, quello che succede è che la proteina possa avere ogni volta la stessa struttura (potrebbe anche non essere così, però l'altro minimo è talmente difficile da raggiungere che è molto poco possibile).

3.1.1 Metodi sperimentali

I metodi per ottenere le strutture a livello atomico sono la cristallografia a raggi x, o la risonanza magnetica nucleare.

Come si fa a ottenere la struttura? Dobbiamo ottenere un cristallo di una proteina, i raggi x saranno diffratti dagli elettroni, il patner di diffrazione ci permette di strudiare gli elettroni.

Facciamo una goccia di proteina più una certa risoluzione, e usiamo lo stesso solvente usato per dissolvere la proteina ma senza la proteina. Aggiungiamo qualcosa, uno ione, un particolare composto nella parte del reservoir, sigilliamo il tutto. Siccome la concentrazione è magiore nel reservuoar, l'acqua evaporerà molto lentamente. In alcuni casi la proteina riesce a formare dei cristalli. Supponendo di aver truvato il cristallo collezioniamo il patern di diffrazione. Siamo ingrado di misurare l'intensità di raggi x diffratti ma non la loro fase, esistono tecniche che si basano sulle trasformate di Fourier che utilizzano un po' di truccehtti matematici riescono ad ottenere la mappa di densità elettronica, conoscendo la sequenza dlela proteina si può cercare di fittare la catena polipeptidica nella densità elettronica. Adesso in alcuni casi riusciamo a farlo in modo relativamente veloce. Data la mappa di densità elettronica il fitting è fatto al computer mentre un tempo era fatto a mano. Altra tecnica è quella della risonanza magnetica nucleare.

Se si prende una soluzione di proteine in un enorme campo magnetico gli spin dei protoni si allineano al campo magnetico. Se mando una radio fraquenza, l'ìdrogeno assorbirà, posso assegnare una frequenza a ciascun idrogeno, supponiamo di conoscere a che frequenza risuonano un prtone, e vedere quali altri protoni modificano la energia di radiofrequenza, quello che otteniamo è un set di distanze tra coppie di protoni.

Abbiamo ottenuto la struttura tridimensionale, questo nel PDB (banca dati Protein Data Bank). Ogni riga in cui ci sono le coordinate inizia con ATOM, un numero consegutivo, l'atomo, l'aminoacido a cui corrispondono, il tipo di catena, il numero di aminoacido, x,, y, z, occupancy e bfactor, nuovamente il tipo di atomo.

L'occupancy va da 0 a 1. Nella mappa di densità elettronica ho visto tutta la densità elettornicoa che mi aspetto? Potrebbe essere che una molecola manca di un atomo, e avrò meno densità. Se la denstià elettronica non è sferica attorno all'atomo vedroò una densità elettronica sferica (di quanto si sta muovendo l'atomo). È sempre una buona idea andare a guardare il bfactor per vedere (sfericità).

I file possono essere visualizzati utilizzando un qualunque programmino.

3.2 Folding delle proteine

Abbiamo visto come sono le proteine una volta che si sono strutturate, come fanno a passare da una struttura disordinata a una struttura ordinata. Sappiamo esattamente la formazione atomica della proteina della struttura ma non la sua geometria finale. Ci si potrebbe chiedere se la proteina si struttura per qualche meccanismo interno della cellula. Tuttavia è stato fatto un esperimento che ci garantisce che la proteina raggiunge la sua struttura indipendentemente se è o meno all'interno della cellula.

Il folding problem è come si passa dalla struttura. Negli anni 60 viene fatto il seguente espreimento, prende una ribonucleasi, che taglia RNA, estrae questa proteina e la mette in una provetta. Controlla che la proteina funzioni, funziona. Questa proteina ha 4 ponti di solfuro, lo sperimentatore riduce i ponti di solfuro, rompe i legami covalenti, e introduce agenti denaturanti. Controlla, la proteina non è attiva. Elimina l'agente denaturante, e riossida per far riformare i ponti di solfuro. Siccome nella provetta c'è solo la proteina, la proteina di suo sa come andare dalla forma denaturata alla forma finale.

È stato foatto un esperimento di controllo. Il primo pezzo dell'espreimento è identico. Inverte nell'ultimo passaggio l'agente denaturante non viene tolto, in questo modo viene prima ossidata la proteina, e poi rimosso l'agente. La proteina non riesce più ad andare nel suo minimo di energia e rimane inattiva.

3.2.1 Paradosso di Levinthal

Assumiamo che i due angoli ϕ e ψ possono assumere solo 2 conformazioni, immaginiamo che la proteina formata da 100 aminoacidi, la proteina possa esplorare una conformazione in 1 femtosecondo.

Ci sono 99 legami peptidici e 198 angoli ψ e ϕ, il tempo richiesto per arrivare esplorare tutto avremo 10^{40} anni.

Le proteine piccole si strutturano nel giro di millisecondi.

Ci si è inventato un modello molto utilizzato che si chama modello HP, che dice: invece di esplorare tutte le possibili conformazioni della proteina costruiamo una griglia. Gli aminoacidi sono solo di due tipi H e P, poi abbiamo che le interazioni favorevoli si hanno tra aminoacidi diversi. Vediamo se riusciamo a trovare l'approssimazione della griglia ad energia minima. Vogliamo trovare un percorso nella griglia per cui gli atomi sono solo nei punti di intersenzione e il percorso che voglio cercare è un percorso self-avoiding. Il numero di possibili soluzioni è enorme. Questo è un problema NP completo (non risolvibile in tempi ragionevole, per cui non si può determinare se può essere risolto in modo polinomiale). Bisogna trovare degli algoritmi approssimati per risolverli. In relatà il numero di configurazioni che si devono valutare realmente è molto minore per svariate ragioni.

Per avere qualche idea di come è fatto questo spazio delle possibili informazione. In una griglia tridimensionale può essere usata per approssimare il disegno delle proteine, ma soprattutto per capire se possono realmente trovare il minimo assoluto.

Abbiamo uno spazio delle conformazioni, e ciascuna conformazione ha un energia, e vogliamo trovare il minimo globale, e vogliamo evitare di finire nella ricerca del minimo locale.

Il primo problema è come si esplora questo spazio enorme? Il secondo è: data una determinata configurazione, come si calcola l'energia? In teoria calcolare correttamente l'energia di una proteina dovrei risolvere l'equazione di Schrödinger per una decina di migliaia di atomi. Quindi occorre trovare un modo per calcolare in tempi sensati una qualche approssimazione della energia della conformazione.

3.3 Algoritmi di minimizzazione

Il nostro obbiettivo è quello di cercare un algoritmo che consenta di capire quale configurazione di angoli permetta di raggiungere il minimo di energia.

Il più banale metodo che possiamo cercare si chiama *steepest descent*: andare nella direzione del gradiente dell'energia, ossia verso dove c'è la massima pendenza:

$$x_{i+1} = x_i + \lambda_i \nabla E(x_i)$$

Poioché il gradiente dell'energia rappresenta la Forza questo algoritmo corrisponde alla seguente sequenza:

- Settiamo una configurazione iniziale
- Calcoliamo l'energia potenziale della configurazione
- Applichiamo a ciascun atomo il gradiente della forza
- Controlliamo se il gradiente è nullo

Se il gradiente dell'energia è nulla abbiamo trovato un minimo locale. È molto difficile che questo sia anche il minimo globale del sistema. Per risolvere il problema ci sono vari trucchi. Un trucco è quello di utilizzare la tecnica del gradiene coniugato. Ci si muove in una direzione ortogonale al quella del gradiente, e poi si prova a minimizzare. In questo modo si resce a superare qualche minimo locale più piccolo. Questo metodo è più probabile che superi il minimo locale, ma è computazionalmente più complesso. Le tecniche più utilizzate sono tecniche invece di tipo probabilistico, come la tecnica Monte Carlo, o algoritmi genetici.

3.3.1 Monte Carlo

Si inizia con una certa conformazione, questa viene perturbata e si calcola l'energia della nuova configurazione. Se la nuova configurazione ha un energia minore me la tengo.

Se la configurazione finale ha un energia maggiore di cui sono partita torno indietro con una probabilità che dipende dalla differenza di energia.

Si calcola il valore $a = e^{-\frac{\Delta E}{KT}}$, estraggo un numero a caso $b \in [0, 1]$, se $b > a$ mi tengo la configurazione, altrimenti torno indietro.

Il senso del metodo Monte Carlo è insito nella statistica di Boltzmann. La probabilità di osservare una certa configurazione è:

$$p_i = \frac{e^{-\frac{E_i}{KT}}}{Z}$$

Quindi la probabilità di passare da una configurazione i ad una configurazione j la possiamo stimare come il rapporto tra p_i e p_j.

La proteina è una molecola. Possiamo modificare le distanze di legame o gli angoli:

- Si sceglie un legame a caso
- Si divide la molecola a metà del legame.
- Scelgo estrarre a caso un numero che mi dice di quanto modificare questo legame.
- Streccio la molecola attorno al legame e mi calcolo la differenza di energia tra prima e dopo
- Test montecarlo se tenere o meno la nuova configurazione.

Per gli angoli la mossa è uguale. Si sceglie a caso un angolo, si divide la molecola attorno all'angolo, si ruota di un valore casuale la molecola attorno a quell'angolo, si ricalcola l'energia e si procede al test montecarlo.

Questo tipo di logica si può applicare a quasi tutti i problemi di ottimizzazione.

I vantaggi del Monte Carlo consistono nel basso costo computazionale, semplicità di focalizzarsi su una regione locale, e la possibilità di fare mosse non fisiche, in modo da superare minimi locali.

Siccome il parametro T è al denominatore, più alto è T più è probabile che io accetti la mossa, fisicamente T rappresenta la temperatura del sistema.

Si può costruire un algoritmo di minimizzazione definendo come deve essere scelta la temperatura.

Il metodo più usato è il Simulated Annealing.

All'inizio della simulazione si parte con un T molto alto, man mano che si procede si abbassa la temperatura. Si chiama simulated annealing perché assomiglia all'annealing delle molecole quando si abbassa la temperatura.

Si può fare poi un altra cosa: *Replica exchange*. Si fanno tante simulazioni in parallelo, in ciascuna delle simulazioni si cambia qualcosa. Ovviamente siccome si tratta di simulazioni casuali tutte le simulazioni non saranno uguali. A certi intervalli si sceglie casualmente due simulazioni e si scambiano tra di loro le configurazioni a cui sono arrivate. E decido se accettare o meno lo scambio in base ad un altro test di monte carlo.

3.3.2 Algoritmi Genetici

C'è un altro metodo molto usato e molto efficiente, che copia quello che succede in natura. Se abbiamo una popolazioni di cromosomi e fanno figli, gli individui più adatti contribuiscono maggiormente alla popolazione finale.

Devo avere una funzione di fitness, per ogni individuo ho un numerino che mi dice qanto è bravo il mio individuo, faccio accoppiare gli individui con fitness migliore, e ottengo una nuova generazione dei figli, nella riproduzione si applica la logica del crossing over e delle mutazioni casuali.

La logica è che non si vuole minimizzare, se si prendono gli individui solo con la fitnes migliore starei ottimizzando io voglio permettere al sistema di esplorare il maggior numero di configurazioni possibili. Permettiamo di esplorare anche dei massimi di energia per trovare anche una ottimizzazione migliore. Cercare di superare i massimi locali.

Il modo più semplice in cui si usano gli algoritmi genetici è di trasformare l'individuo in una stringa, si deve codificare la nostra proteina in una qualche stringa. La popolazione inizale saranno tante stringhe, che rappresentano un gran numero di possibili conformazioni. Scelgliamo a caso un individuo e faccio una mutazione. Seleziono due individui e gli faccio fare la progenie, si prende la prima metà di un individuo e la seconda metà di un altro, si possono prendere dei pezzi diversi, questo dipende dall'algoritmo. Si ottiene una nuova popolazione, si valuta la fitness dei nuovi individui e si ricostruisce una popolazione inizale che valuta la fitness dei figli. Man mano che si va avanti, la popolazione si arricchirà di individui con fitness migliore.

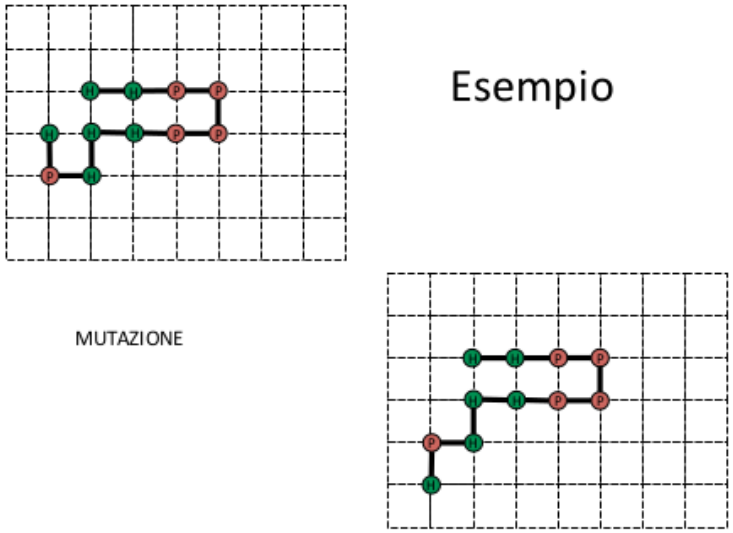


Figura 3.4: Esempio di una possibile mutazione in un algoritmo genetico.

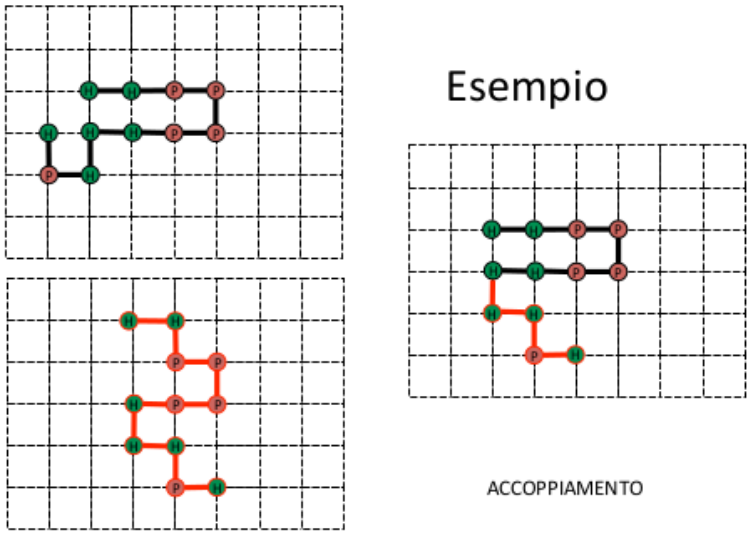


Figura 3.5: Esempio di un possibile Cross Over tra due individui.

3.4 Calcolo dell'energia

Per fare tutto questo ci serve la nostra funzione di fitness che è l'energia. Come si calcola l'energia della proteina? In linea teorica bisognerebbe risolvere l'equazione di Schrödinger, che è un problema difficilissimo. Si devono fare delle approssimazioni. Ansitutto ci si mette in regime classico; Le forze che determinano la struttura di una proteine sono i legami covalenti, gli angoli fra gli atomi, gli anoli diedri, ecc. Dobbiamo calcolare l'energia data la struttura.

I legami sono approssimati come una molla. Esiste una certa distanza ottimale tra i due atomi r_0, e la molla ha costante k. La forza con cui sono tenuti assieme di un legame doppio è maggiore della forza che tiene assieme un carbonio ed un idrogeno.

Per sapere le forze in gioco si possono prendere gas noti (come il metano) e calcolarle.

Questo insieme di parametri si chiamano campi di forze o *force field*. Ce ne sono svariati in letteratura. Se si vuole calcolare l'energia di una proteina occorre riconoscere il tipo di atomi e il timo di legame. Anche l'angolo tra tre atomi si approssima come una molla. Di nuovo a seconda del tipo di atomi e di legami, avremo θ_0 e k.

La stessa cosa vale nell'angolo diedro, in cui però usiamo una funzione periodica del tipo:

$$E = k_\varphi \left[1 + \cos(n\varphi + \delta)\right]$$

Poi abbiamo le interazioni tra cariche, che studiamo con il potenziale di Coulomb. In relatà bisogna considerare lo schermaggio delle cariche per opera delle molecole d'acqua. Questo lo si fa aggiungendo molecole d'acqua alla simulazione (non usando la costante dielettrica che è macroscopica). Poi ci sono le forze di vander walls, per cui si usa il potenziale di tipo $r^{-12} - r^{-6}$. Da cui l'energia diventa:

$$E_c = \sum_{b=\text{legami}} k_b(b_c - b_0)^2 + \sum_{\theta=\text{angoli}} k_\theta(\theta - \theta_0)^2 + \\ + \sum_{\varphi=\text{Angoli diedri}} \frac{K_\varphi}{2}\left[1 + \cos(n\varphi - \gamma)\right] + \sum_{i,j}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon_0 |\vec{r}_i - \vec{r}_j|}\right]$$

In questo modo siamo in grado di calcolare l'energia di un certo sistema, a patto di avere i parametri per tutti gli atomi del sistema.

Se nella nostra proteina c'è un atomo strano, non ci sono i parametri per quell'atomo e li diventa un campo di ricerca. In pratica data la posizione di tutti gli atomi di una proteina si può calcolare questa energia approssimata come la somma su tutti i legami di quest parametri, un termie sui legami un termine sugli angoli uno sulgli angoli diedri più una somma su tutti i parametri.

Ogni tatno si aggiungono termini di correzioni. Data una proteina siamo in grado di calcolare la forza approssimata che agisce su ogni atomo, ci sono una serie di discussioni in giro. Si prende una sequenza aminoacidica di una proteina, si riesce ad arrivare alla struttura nota della proteina? La risposta è no.

Nessun metodo oggi è in grado di farlo. Questi metodi si chiamano ab initio. Perché questo non si riesce a fare? Perché sto calcolamdo male l'energia o perché non funziona bene monte carlo?

Una cosa divertenti è che un signore ha fatto un computer che è in grado solo di poter fare le simulazioni delle proteine, il risultato di questa cosa è che all'inizio sembrava che arrivava alla struttura delle proteine, ultimamente sembra che invece ci riesce raramente.

Un altro esempio di supercalcolo che si può provare a fare è una cosa che si chiama folding at home. Esistono dei problemi che sono parallelizzabili. Algoritmi che si possono far girare tanti computer. A ogni computer si parte da una posizione diversa, era partito questo progetto che si chiamava SETI@home. Questo da un programmino, che il computer quando non sta facendo nulla continua una simulazione montecarlo, che a cui a ciascuno viene data una conformazione iniziale di una proteina, dopo raccoglie il risultato migliore. È come stare usando il computer più potente dell'universo, questo sembra funzionare per proteine molto piccole. È estremamente fastidioso non riuscire a risolvere il problema usando solo basi fisiche.

3.4.1 Dinamica Molecolare

Se si conosce l'energia se si conosce la forza forse si può risolvere l'enquazione di Newton. Le posizioni inziiali le possiamo ricavare. Le velocità iniziali le possiamo prendere dall'equazione di Boltzmann, tale per cui la distribuzione di velocità sia uguale a quella di Boltzmann. La forza ce l'abbiamo, possiamo vedere come si spostano gli atomi nel tempo delle proteine. Questa è l'idea di base, si riesce a calcoalre la forza di ogni atomo, se il Δt di interazione delle equazioni del moto è piccolo si può vedere come si muove la proteina.

La dinamica molecolare serve ad un sacco di cose, si può usare per esplorare lo spazio delle conformazioni. In qualche modo si può vedere la proteina che si muove nel suo spazio delle conformazioni. La forza è vero che è la derivata dell'energia, ma si possono applicare altre forze a questi atomi.

Se Si possono aggiungere aggiunge altre forze per rendere il problema più sempolice, ad esempio forze elastiche tra aminoacidi lontani se so che nella struttura ripiegata in realtà sono vicini. Questa dinamica molecolare mi permette di aggiungere dei vincoli e fare in modo di tener conto sia della parte chimica della proteina, aggiungendo dei vincoli che dipendono dall'esperimento e dal problema. Il problema della dinamica molecolare sorge sulle somme tra tutte le coppie di atomi. Quando si pensa al calcolo dell'energia c'è anche l'energia di tutte le molecole d'acqua che sono intorno, il numero di atomi del sistema è molto alto. Nelle simulazione più lunghe si riesce ad arrivare nel giro di millisecondi.

Come facciamo a schematizzare l'effetto di una proteina in un solvente? Quello che si fa è che si sceglie un cubo che possa contenere la proteina e la molecola di acqua e poi si utilizzano le periodic boundary conditions. Questo da ulteriore complessità computazionale. La dinamica molecolare può essere usata in modo molto più interessante. Non va bene neanche mettere solo acqua pura, ci sono simulazioni fatte all'interno delle membrane che invece dell'acqua formano un doppio strato lipidico. Io qui sto seguendo il comportamento di ciascun atomo dato le forze che sono in gioco. L'altro problema della dinamica molecolare è che con questo metodo sto calcolando l'energia ma non l'energia libera, se l'entropia aumenta o diminusce non ne sto tenendo minimamente conto.

In generale si tende a dire che le interazioni tra coppie di atomi sono limitate entro un certo raggio, decido una distanza entro la quale solo quelli all'interno di un certo raggio influenzano il comportamento. Si risolve l'equazione del moto in un intervallo piccole.

C'è un problema che durante la simulazione dobbiamo garantirci che i parametri macroscopici sia senzati, la temperatura deve essere costante e anche la pressione, si fa il passo seguente e si continua questo circolo.

Anche in queste simulazioni si deve includere la temperatura e può essere fatta anche qui la simulated annealing. Si parte da temperature più alte, e man mano abbassare la temperatura. Chiaramente si può immaginare di fare una simulazione irraggionevole se lo scopo è quello di esplorare un numero sempre maggiore delle conformazioni.

3.4.2 Teoria del Funnel

La teoria corrente che non risolve completamente il paradosso di Levingein. Quando la proteina è nella sua forma destrutturata, ha una enorme guadagno di entropia e entalpia, perché aumentano le interazioni di Van der Waals. Esistono proteine che si strutturano seguendo sempre lo stesso percorso. Una teoria è quella che ci dice che nello spazio delle conformazioni esiste un folding pathway, una volta che la proteina raggiunge questo percorso per caso collassa verso lo stato di minimo funzionale.

Esiste un altra teoria detta Teoria del Funnel: Un contatto tra due aminoacidi molto favorevole,dDurante il processo di folding, rimane più a lungo rispetto agli altri. Con questa ipotesi sbaglia il tempo di calcolo della formazione della proteina di qualche ordine di grandezza. Quello che la cellula deve assolutamente evitare è che la proteina rimanga troppo a lungo in un minimo locale. In una strutturazione sbagliata se avessi regione idrofobiche troppo esposte potrei avere degli aggregati che sono insolubili in acqua. Si formerebe un grumo di proteine, è quello che succede nell'alzaimer e nella SLA. L'energia termica deve essere sufficiente a destrutturarla e ristrutturarla in un altro modo. Perché questo possa succedere le barriere di potenziale tra i minimi locali non possono essere così alte. Affinche tutte le barriere siano piccole, basta ipotizzare che anche i minimi locali siano piccole.

3.4.3 Energia euristica e potenziali di coppia

Possiamo anche fare una stima più grossolana dell'energia in modo da esplorare più rapidamente lo spazio degli stati, e poi usare funzioni più complete solo quando ci troviamo vicino alla regione del minimo assoluto. Approssimiamo ogni aminoacido con un punto. Possiamo date le strutture note calcolare la frequenza con cui gli aminoacidi si trovano in una certa regione.

Ignoriamo i dettagli della sua catena laterale, una coppia di aminoacidi quanto frequentemente li troviamo vicini ad una certa distanza? Questa frequenza si può approssimare ad una probabilità. Grazie all'equazione di Boltzmann abbiamo una corrispondenza tra probabilità ed energia.

$$p = e^{-E/Kt}/Z$$

Si calcola la probabilità di avere aminoacidi ad una certa distanza, facendo il rapporto di probabilità tra aminoacidi diversi si ottiene la somma o la differenza delle energie.

Possiamo anche calcolare qual è la frequenza con cui mi aspetto di trovare per caso due aminoacidi a quella distanza, se la probabilità empirica è migliore allora l'energia è minore.

Presno ad ogni aminoacido si calcola tutto per una coppia e per la coppia successiva. Si calcolano le distribuzioni per ogni tipologia di distanza, per ogni distanza dello spazio e lungo la catena abbiamo una frequenza. Che se dividiamo per la distanza casuale ci dice se questa è una distribuzione casuale o meno. Tutto questo conto lo faccio su tutte le proteine della banca dati. Per ottenere la distribuzione casuale si inverte casualmente gli aminoacidi, e ci ricalcoliamo le distribuzioni che a questo punto sono casuali.

Questi sono detti potenziali di coppia. In pratica sono delle tabelle che per ogni coppia di aminoacidi ci dicono l'intervallo di separazione, d i quanto sono separate lungo la sequenza, e l'energia. Questi potenziale di coppia tipicamente si usano prima, e poi si va sempre più a potenziali meno approssimati.

L'altra è la valutazione del punteggio dei vari modelli.

3.4.4 Proteine omologhe

Possiamo sfruttare proteine omologhe: È l'evoluzione che ha scelto tra tutti i possibili polimeri di aminoacidi finissero spontaneamente sempre più o meno nella stessa struttura. La stabilità è marginale. La probabilità che prendo una struttura a caso degli aminoacidi ed è una proteina è bassissimo.

Supponiamo di avere una mutazione di un aminoacido nella proteina. Probabilmente la consequenza sarà che la proteina non si struttura più. L'altra probabilità è che faccia un'altra struttura. L'altra possibilità che ci interessa è che la mutazione si accomoda nella struttura della proteina. Queste mutazioni sono riuscite a sistemarsi all'intero della struttura, non da proprio fastidio. Il risultato di questo ragionamento è il seguente. Proteine omologhe hanno una struttura tridimensionale simile, perché le mutazioni se son state accettate è perché sono state in grado di accomodarsi all'interno della struttura con dimensione locale. Se le proteine omologhe hanno struttura simile abbiamo un jolly. Se abbiamo risolto la struttura dell'emoglobina del cavallo, ho un idea di come è fatta l'emoglobina dell'uomo, del cane, del maiale... Avere informazioni su una proteina della famiglia mi da informazioni sulla struttura di tutta la famiglia.

Come si misura la similarità tra due proteine? Vorrei poterle sovrapporre (ottimizzo la sovrapposizione) e vedere di quanto distano gli atomi corrispondenti. Per poter sistemare la stessa distanza, si può usare la backbone, le catene principali. Le sovrappongo in modo di minimizzare la distanza e poi misuriamo la distanza, di solito il valore che si usa è RMSD.

Prendiamo coppie di proteine omologhe? se mettiamo sull'asse delle x c'è la percentuale di ugualgiamza, proteine identiche non hanno la stessa distanza è dovuta al fatto che sono risolte da laboratori di versi. Man mano che le proteine hanno uan minore identità di sequenza anche la loro differenza strutturale aumenta. Due proteine con il 50 % di aminoacidi identici, in media si differenziano di 10^{-10} m (le posizioni medie degli atomi).

Si cerca la proteina della stessa famiglia evolutiva. Se tra le proteine omologhe ce ne è una di struttura omologa, questa mi formisce una approssimazione della prima struttura. Immaginiamo di avere due proteine strutturate. Se abbiamo l'allineamento tra le due sequenze, possiamo dire, si può ricostruire in qualche modo la struttura, si parte dall'allineamento e si copiano le coordinate degli aminoacidi corrispondenti. Questo va bene fin quando le regioni sono conservate. Se questa proteina è quella di struttura nota. Ottengo le coordinate delle insersioni conservate. La cosa fondamentale di un allineamento è quello di averlo corretto. Abbiamo discusso che gli allineamenti multipli sono più affidabili degli allineamenti di coppia.

Il problema di modellare regioni che hanno cambiato conformazioni, di quelle con insersioni e delezioni sappiamo che hanno cambiato forma. In genere si possono tenere nella proteina. Regioni che non siano elementi di struttura secondaria. Nella maggior parte dei casi le regioni strutturalmente divergenti corrispondono alle anse e alle upse. Spesso viene detto il problema della predizione dei Loop (perché sono quelle in cui si verificano insersioni e delezioni).

Quando abbiamo costruito la struttura della catena principale per la maggiorparte delle regioni, si può provare a inserire tutti i possibli modi con il possibile numero di aminoacidi. Qualche caso si può fare, però abbiamo un altra possibilità. Supponiamo di dover inserire 5 aminoacidi. Si possono costruire tutti i possibili modi in cui la catena può andare tra i due capi. Tra le proteine note, in generale per coprire quella distanza con quel numero di aminoacidi che fanno? Si può andare nella banca dati e tirare fuori tutti i loop che hanno la distanza e il nuero giusto di aminoacidi che sono quelli della sequenza, e se ne sceglie 1.

Abbiamo costruito il modello della proteina nel modello backbone. Si vanno a prendere frammenti di altre proteine e si sceglie quella ad energia minore. Si otteiene un modello di una proteina non proprio uguale alla proteina omologa da cui ho iniziato. La proteina omologa la chiamiamo "template" e quella che vogliamo modellare si chiama "target". In alcuni casi le catene laterali saranno uguali. Qualche catena laterale va ricostruita. Questo è un altro problema np completo. Se si volessero esplorare tutte le possibili per le catene laterali siamo punto e accapo. Quando non abbiamo sufficiente forza computazionale si va a vedere cosa è successo in natura.

Nelle banche dati osserviamo un numero limitato di conformazione, ciascuna delle quali ha una sua frequenza. Nuovamente possiamo collegare attraverso la legge di Boltzmann la probabilità all'energia. Anziche esplorare tutte le possibili conformazioni posso esplorare solo le conformazioni che ho già osservato. Questa è una specie di alberlo decisionale. Una delle tecniche di ottimizzazione che abbiamo visto, spesso si usa il dead end, che trova l'albero con costo minore. Anime buone hanno tabulato la frequenza con cui si osservano le conformazioni delle varie proteine. Quello he facciamo è esplorare solo le conformazioni delle librerie di rotameri. È sempre una buona idea se si possa esprimere il problema sotto forma di albero si possono usare una serie di algoritmi ben noti per usare gli alberi.

Se troviamo più di una proteina omologa possiamo valutare le distribuzioni di probabilità di angoli e distanze degli aminoacidi in funzione del tempo evolutivo, e quindi estrapolare da questi dati dei vincoli nuovi su cui minimizzare la proteina (dalle probabilità si passa all'energia con l'equazione di Boltzmann).

Aggiunti dei vincoli riducono lo spazio da esplorare. Questo metodo è quello che funziona meglio di tutti.

3.4.5 Metodi su frammenti

I metodi per la ricerca di analogie basati su modelli nascosti di Markov hanno aumentato tantissimo i database di struttura omologa. Supponiamo di prendere una proteina, i primi 8 amionacidi, vediamo se troviamo anche gli stessi 8 aminoacidi hanno la stessa struttura in due proteina sequenza, non molto spesso trovate che abbiano la stessa

conformazione. Prendiamo la sequenza e la tagliamo a pezzi non sovrapposti, e si ricerca questa sequenza in una poteina di struttura nota, salvo che in ogni proteina avrà una struttura diversa. Avremo per ogni frammento una collezione di varie possibili conformazione. Dopo di che si combinano in maniera casuale i vari frammenti. Si costruisce tutti i modelli. Ho una popolazione iniziale.

Simulated annealing o algoritmi genetici possono evolvere questa popolazione iniziale.

Gli algoritmi di questo tipo si diferiscono per dimensioni di frammenti, metodi di ottimizzazione e funzione energia (Force fields).

Si possono usare anche dati prevenienti da altre predizione, sappiamo con metodi di apprendimento automatico predire la struttura secondaria, si può selezionare questo per selezionare i frammenti. Ci sono metodi sempre analoghi per predire se l'aminoacido è esposto o all'interno. Una cosa che non si capisce bene se funziona o no è la predizione dei contatti.

Se si ha un allineamento multiplo di proteine, possiamo chiederci se esistono variazioni correlate. Se due variazioni di aminoacidi sono correlate in tutto un allineamento probabile che quegli amnoacidi siano vicini (ma potrebbe non funzionare bene).

Come facciamo a sapere se un metodo di predizione funziona o no?

Hanno inventato un esperimento che si chiama CASP. Prima che chiunque abbia risolto sperimentalmente la struttura sperimentale senza risolverla? Ci sono dei valutatori. Questo era nato per valotare i metodi di risoluzione delle strutture ora si è espanso.

3.5 Progetto CASP

Il progetto CASP (Critical Assessment of techniques for Proteine Structure Prediction) serve per valutare i metodi di strutturazione delle proteine, dobbiamo cercare di capire quali sono le parti dell'intero processo che richiedono ulteriori approfondimenti.

Storicamente era dedicato solo alla predizione di struttura tridimensionale e secondaria, le predizione di struttura secondaria sono diventate talmente accurate che ha perso interesse, nel frattempo si sono aggiunte altre categorie, modellizzazione in cui i contatti tra aminoacidi siano noti o predetti, analisi della qualità dei modelli (ti do N modelli e me li metti in ordine di accuratezza).

C'era anche una categoria per predire le regioni di proteine che sono disordinate, si sono scoperte l'esistenza di proteine che hanno regioni non strutturate, si strutturano nel momento in cui incontrano il patner di strutturazione. Ci sono una serie di teorie, immaginiamo di avere due proteine che interagiscono che abbiano una grande specificità, e contemporaneamente non hanno una grande affinità di interazione (possono staccarsi facilmenti). Per ottenere una grande specificità la cosa migliore è che le proteine facciano tante interazioni, grande superficie di interazione corrisponde ad una grande affinità. Questa è compensata dall'energia che viene persa quando si struttura la proteina.

I risultati di Casp sono abbastanza rappresentativi di quello che c'è nel mondo, ci sono circa 350 gruppi che partecipano, e il numero di predizione che arrivano sono decine di migliaia. Tutte queste migliaia di predizioni vanno analizzate in modo molto veloce. Esiste un sito dove vengono messi tutti i modelli con delle valutazioni.

Funziona nel seguente modo:

- Raccolta delle sequenze di proteine che stanno per essere risolte in laboratorio.
- Raccolta delle predizioni per ciascun target.
- Valutazione dell'accuratezza del metodo anonimamente.

Se però abbiamo la struttura della proteina e il modello, vogliamo sovrapporre la proteina e il modello in modo da ottenere una stima di che errore è stato fatto.

Ci sono vari modi per valutare questo fit. Uno è la distanza media tra gli atomi corrispondenti. Nel caso di predizioni di struttura questo non è il modo più geniale per confrontare le strutture. In genere si usano altri parametri che mi dicono ad esempio la distanza media tra tutti gli atomi, non peso quadraticamente ciascun errore. Questa misura si chiama GBTPS.

Perché abbiamo bisogno dei valutatori? Ogni linea del grafico rappresenta una predizione per un certo target, la linea rossa è stata modellata la il 40 % della struttura a meno di un amstrong. Cioè un modello ha ottimizzato meglio la distanza sotto al 40 %, mentre un altro può averla analizzata meglio in generale (linea blu), ma con meno precisione (Figura 3.6).

La forza di questa valutazione risiede nel fatto che tutti i metodi sono valutati sulle stesse proteine. Come si comportano i server automatici? Se un giorno abbiamo bisogno di partire da un modello ragionevole della proteina conviene inviare la nostra sequenza di aminoacidi ad uno di questi server. HhpredB ad esempio ricerca gli omologhi con Markov, e poi usa Modeler per costruire la proteina.

Nel caso del Free modelling è più difficile valutare i modelli. Uno dei problemi che è se possediamo più modelli delle proteine è molto probabile che tra di loro ci sia un modello che si avvicina molto alla proteina finale, ma sembra

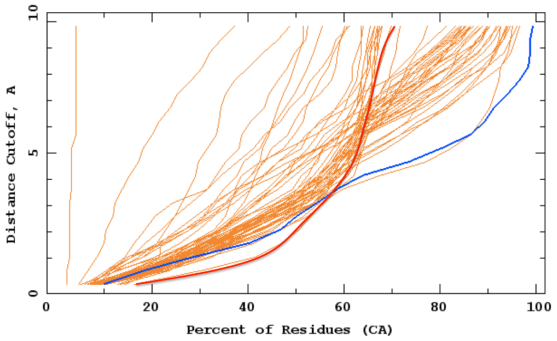


Figura 3.6: Percentuale dei residui dei metodi in CASP.

che non siamo in grado di dire qual è. Il problema di selezionare correttamente è un problema buffo, se si prendono questi modelli.

L'altro problema è che bisogna fare attenzione alla biologia della proteina (se ci sono strutture quaterniaria).

Se abbiamo un modello approssimativo della proteina, riusciamo a rifinirlo in modo da averne uno più preciso? Ancora no. Questo problema c'è speranza di migliorarlo dal punto di vista fisico. Le regioni strutturalmente divergenti non le sappiamo modellare bene.

Un altro problema interessante è capire la funzione delle proteine. Se abbiamo una sequenza di proteine, che funzione hanno? Anche in questo caso ci sono una serie di esperimenti che seguono CASP (Come ad esempio CAFA). Si prende un enorme numero di sequenze e si mettono a disposizione, quelli che partecipano predicono una funzione per queste proteine, dopo un certo periodo di tempo qualcuna di queste proteine sarà stata valutata in laboratorio e su queste si valuta la qualità. Come si definisce la funzione della proteina è la gene ontology.

La gene ontology è un grafo diretto aciclico in cui ogni nodo dell'albero può avere più di un genitore. L'ontologia che assegna le funzione delle proteina è formata da grafi aciclici. La funzione può essere descritta in modi differenti, la proteina appartiene a quel gruppo. La trombina fa parte del processo di coagulazione del sangue, la trombina taglia il fibrigenogeno quindi è una proteasi (taglia un legame polipeptigono), è extracellulare. Bisogna valutare quanto bene è stato predetto il processo biologico, la struttura di questo grafo è molto complessa. Dipende da quanta informazione viene predetta.

3.5.1 Interazioni macromolecolari

La maggior parte delle proteine non funziona da sola, ma funziona perché interagisce con qualcosaltro, gli anticorpi devono riconoscere altre proteine o zuccheri, o acidi nucleici. La differenza di energia tra stato complessato è negativo rispetto a quello libero. Data la struttura di due proteine la domanda è: Interagiscono?

Esiste qualche metodo che può utilizzare le seguenze. Possiamo fare una cosa banale. Abbiamo tanti genomi, poi abbiamo delle proteine, in quali genomi quella particolare proteina è presente, sappiamo identificare gli omologhi o meno, se c'è metto 1 se non c'è metto 0. Dopo di che se ci sono delle proteine che sono sempre assieme o quando non c'è l'una non c'è neanche l'altra posso dedurre che quelle proteine hanno qualcosa a che fare. Se in un genoma abbiamo due proteine, e in un terzo genoma ne troviamo una che è la fusione delle due, allore è molto probabile che quelle due interagiscono l'una con l'altra. L'ultimo metodo è uno che è stato già descritto. Possiamo guardare le mutazioni correlate. se due proteine mutano insieme possiamo fare l'ipotesi che queste due proteine siano vicine nello spazio, le loro proprietà devono essere complementari, quando uno cambia, cambia anche l'altro. Questo può essere vero anche nelle regioni di approssimazione. Prendiamo la proteina A si cercano tutti gli omologhi, si prende la proteina B e gli omologhi. Poi si estraggono coppie di omologhi delle stesse specie, e cerchiamo le mutazioni correlate.

Un metodo che funziona un po' meglio è quello del mirror tree. Possiamo calcolare degli alberi basati sulla sequenza di una certa proteina. Quando si vogliono fare queste cose per capire meglio si deve lavorare su regioni funzionali. Ogni proteina ha una sua velocità di evoluzione, se prendiamo proteine diverse e costruiamo alberi filogenetici, alberi simili corrispondono a proteine che evolvono similarmente, e quindi posso supporre che sia accurato. Se noi facciamo un prescreening con questi metodi allora abbiamo molto meno candidati che interagiscono. Se sappiamo che due proteine interagiscono, vogliamo sapere come, i metodi che cercano di risolvere questo problema sono i metodi di docking. Metodi analoghi sono utilizzati, dati una proteina e una piccola molecola, dove interagiscono? Se abbiamo due proteine e vogliamo trovare come interagiscono, dovremo esplorare trutte le traslazione e rotazioni di una rispetto all'altra, e scegliere la conformazione che rende minima la differenza di energia libera.

Per quello che riguarda le molecole biologiche, è molto semplice descrivere il problema nella maniera ideale, impossibile nel mondo reale. Se anche qualche atomo di una proteina è vicino all'atomo dell'altro, faccio finta di nulla, perché è possibile che quegli atomi che non hanno una interazione favorevole si spostino. Oppure si possono prendere

le due proteine, faccio una simulazione di dinamica molecoalre, prendiamo le conformazioni più frequenti e faccio il Docking utilizzando queste conformazioni. L'altro metodo è quello dell'esplorazione di tutte le confomazioni.

Abbiamo due proteine dobbiamo sapere dove interagiscono, di solito si chiamano recettori e ligando. Prendiamo la proteina e mettiamola in una griglia, all'interno della griglia segno un valore ad ogni cella della griglia, tutti quelli in cui la proteina non c'è hanno un valore zero. Tutti quelli che sono dentro gli metto un valore grande, stessa cosa con il ligando. Se la superficie del ligando fitta esattamente quella del recettore sarà 1, sennò 0 altrimenti enorme. Posso immaginare con operazioni di rotazioni e trasformazioni e mi calcolo il prodotto cella per cella della prima e della seconda, e sommo. Calcolare questo oggetto ci vuole un sacco di tempo. Possiamo usare le trasformate di Fourier. Usando le trasformate di Fourier cresce come $N^3 \log N^3$ ansiché di N^6.

$$S(R,T) = \sum_{i,j,k=1}^{N} a(i,j,k)b'(i+T_x, j+T_y k+T_z) \qquad \hat{S} = \hat{A}^* \cdot \hat{B}$$

Dove con $\hat{S}$ abbiamo indicato le trasformate di Fourier, in questo modo si guadagna parecchio in fatto di rapidità computazionale.

Per identificare le regioni concave o convesse si usano due metodi. Si rappresentano la superficie della prooteina con delle sfere, con dei cerchio...

Ci sono dei cambi conformazionali quando due proteine interagiscono, come esploro le possibili conformazioni e come misuraimo la loro energia di ciascuna informazione?

Rosetta dock parte da una posizione di partenza, si prendono un po' di orientamenti casuali, si fa una ricerca montecarlo a bassa risoluzione, per filtrare le informazioni, partendo dalla conformazione casuale, si fa una perturbazione, si valuta Monte Carlo, e si utilizza come energia i potenziali di coppia.

Abbiamo il potenziale di coppia tra i due residui, hanno un energia più negativa se gli aminoacidi sono quelli sulla superficie, si possono avere delle indicazioni che questo aminoacido devono stare da una parte dell'interfaccia. A questo punto si fa un ottimizzazione più dettagliata, perturbazioni casuali, ottimizzazione delle catene laterali, minimizzazione a corpo rigido. Si possono usare una serie di filtri. Si possono mettere filtri a qualunque livello, sia di informazione biologica (bassa risoluzione) che basati sull'energia (alta risoluzione).

Alla fine del Monte Carlo otteniamo più soluzioni. Una volta ottenuto tutti i possibili orientamente facciamo un clustering confrontando a coppie le soluzioni. Se ci sono più regioni in cui si sono concentrate più soluzioni, più sono popolati quei gruppi più è probabile che quella sia la soluzione corretta.

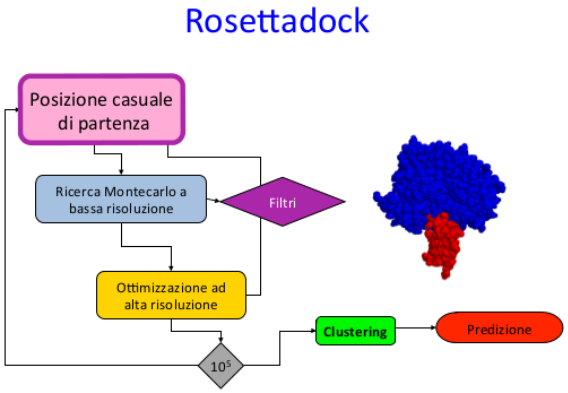


Figura 3.7: Schema di funzionamento di Rossetta Dock

Funzionano questi metodi? Prendiamo il caso in cui si riconosce i risultati e confrontiamo con le previsioni computazionali: esiste un altro esperimento che si chiama CAPRI che fa esattamente la stessa cosa di CASP.

Esistono metodi sperimentali: Il cross-linking vuol dire che abbiamo le proteine, si usano agenti chimici, si tagliano le proteine a pezzi, riusciamo a capire quale frammenti sono stati legati dall'agente chimico. Uso NMR per riconoscere gli scambi di Deuterio idrogeno (una proteina la sintetizzo con Deuterio, le altre con l'idrogeno, via NMR posso distinguerli e tramite spettroscopia posso capire quali muclei hanno cambiato livello energetico). I vincoli possiamo ottenerli attraverso semplici esperimenti. Una delle cose gradevoli è che i vincoli si possono utilizzare il prima possibili, e possono essere utilizzate per filtrare le soluzioni a tutti i livelli del processo.

Capitolo 4

Network Biologici

L'altro tipo fondamentale di dati che si ottengono nel mondo biologico sono le reti, La rete è composta da nodi o vertici, connessi tra loro (edge). Dal punto di vista biologico abbiamo una serie di network di cui disponiamo:

- Co-espressione
- Trascrizionali
- Interazione tra proteine
- Metabolici
- Gene- Patologia

Se mettiamo tutte le proteine in una rete, alcune proteine si legano a monte di geni di altre e ne permettono l'espressione. Abbiamo reti fisiche, che ci dicono come interagiscono le proteine, reti metaboliche, il metabolita A viene trasformato in B da un emzima (proteina che agiscono da catalizzatore). Anche questo è rappresentato da reti. Poi abbiamo una serie di associazioni tra geni e patologie, connettere geni che sono associate alla stessa patologia o connettere patologie che sono associate alla stessa genomica. Si possono decidere quali sono le connessioni e quali i nodi del problema.

Si possono usare reti per classificare formazioni biologiche.

Dobbiamo dare qualche definizione, il grado di un nodo è il numero dei nodi con cui è connesso. La degree distribution $p(k)$ è frequenza che un certo nodo abbia k connessioni. Il grado medio dei nodi è:

$$d = \sum_{k \geq 0} kP(k)$$

Esistono sia reti non direzionali, che reti direzionali. In generale si troano quattro tipi di rete, la rete casuale, regolare, small world e scale free. Questa classificazione ha a che vedere con il grado dei nodi della rete. Una rete può essere trasformata immediatamente in una matrice. Sulle righe e colonne si mettono tutti i nodi e gli elementi i,j è quante connessione ci sono tra i esimo e j-esimo nodo.

La rete scale free è che in effetti è una specie di frattale. Qualunque regione della rete ha le stesse caratteristiche dell'intera rete. Se guardiamo le matrici di connessione appaiono come in (Figura **??**)

Nella scale free ci sono pochi nodi con molte connessioni e molti nodi con pochi connessione. Una rete scale free è più robusta rispetto ad attacchi casuale, se rompo una connessione, ho un alta probabilità di distruggere un nodo poco importante (sono però sensibile ad attacchi mirati). La random network ha una $P(k)$ binomiale:

$$P(k) = \frac{e^{-d} d^k}{k!}$$

Il network scale-free:

$$P(k) \propto k^{-c}$$

L'esponente di k definisce la topologia della rete. Se guardiamo dati che provengono esperimenti di interazioni proteina proteina, anche questa è più o meno scale free, tipicamente le reti veri tendono a deviare per k alto (c'è un numero limitato di nodi). Anche nelle reti metabolico si hanno reti direzionale, in cui tutti e due sono più o meno scale free.

Spiegare il preché di questo può non essere banale. ù Queste proteine che sono degli hub, dei nodi con cui interagiscono tutte, hanno caratteristiche particolari. Un altra cosa che si può estrarre dalle reti è il coefficiente di clustering: la percentuale di coppie di nodi che sono connessi a V e tra di loro rispetto al numero totale di vicini. Tutti i nodi che con V formano triangoli sono cluster. Ovviamente questo serve per capire quale network è più compatto. Il coefficiente di Clustering dell'intera rete è definita come la media dei coefficienti di clustering su ciascun nodo. Una

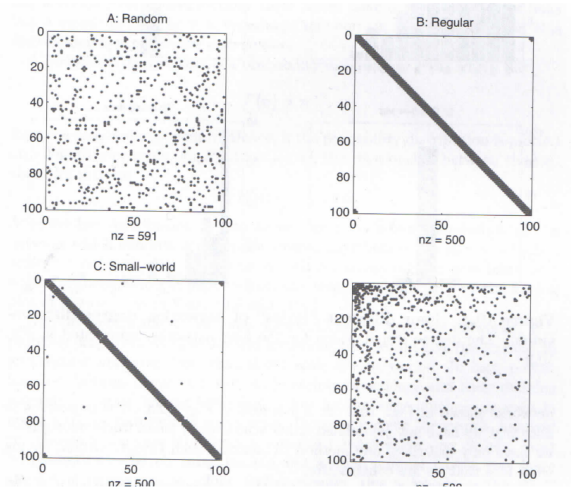


Figura 4.1: Matrici di connessione.**??**

rete completamente connessa ha coeficiente di clustering massimo. Quando abbiamo un sottoinsieme della rete con coefficiente di clustering pari a 1 si chiama clique, queste sono regioni particolarmente compatte che presumibilmente ci danno informazioni molto importanti.

Internet è particolarmente connessa, il 24 % dei nodi di internet è completamente connesso. I siti del www hanno anche essi un coefficiente di clustering molto alto. Anche la catena elementare è fortemente connessa. Possiamo chiederci qual è la distanza media tra due punti della rete? Quanti nodi devo attraversare per andare da un nodo all'altro. Posso calcolare il numero di percorsi che passando da v sono connessi. Componente gigante, dimensione della più grande componente connessa.

C'è un problema aperto sulle reti, non abbiamo una misura che ci convinca sulla robustezza della rete. Non abbiamo una formula analitica per calcolare la robustezza della rete.

All'interno della rete si possono trovare dei motivi, delle topologie di connessione che sono ripetute più spesso di quello che ci aspettiamo per caso. Questi TF sono geni i cui prodotti controllano l'espressione degli altri geni. Esiste un gene z che è' controllato da un trascrittore y che è controllato da x. Poi x a sua volta controlla anche z. Questo modulo è utile per evitare errori di trascrizione, ed è un motivo del genoma. Questo è un modo per eliminare il rumore e reazioni spurie, molto spesso questo tipo di circuiti vengono modellati sotto forma di circuiti elettronici.

Si possono analizzare le reti, biologiche e non, per quali sono i motivi più frequenti. Come si trovano dei motivi in una rete? Si generano un gran numeri di grafi casuali, e si vede se i motivi che cerco nella rete. Occorre generare dei grafi random che rispettino la maggior parte delle propietà che sto analizzando salvo l'identità dei nodi. Questi devono rispettare la topologia. Il motivo che stiamo analizzando è ritrovato ugualmente frequentemente nelle distribuzioni casuali rispetto a quelle reali. Se si vuole generare un network scale free, ci si calcola il valore della costante c, e poi si simulano reti casuali che abbiano la stessa distribuzione, cercando di fare in modo di rispettare la stessa distribuzione di probabilità della rete originare. La maggior parte delle reti che osserviamo.

Dato un network quindi oltre che ad analizzarlo si possono cercare se all'interno di quel particolare network ci sono delle topologie particolarmente rilevanti. Si valuta la "non casualità" del mio motivo con lo Z score, numero medio di volte in cui lo vedo nella mia distribuzione casuale diviso la deviazione standard. Più è alto Z vuol dire che quella particolare topologia è osservata più spesso di quello che mi aspetto per caso. Ci sono vari moduli che sono stati analizzati.

Tutte questi definizione servono a caratterizzare una particolare rete per capire che tipo di rete è, il fatto che la rete di interazioni tra proteine sia scale free ci da informaizoni sulle proteine, se si guardano i patner dei fattori di trascrizione vediamo che ci sono alcuni potenti che agiscono su un enorme numero di geni, e tantissimi specifici che agisocno su pochi geni. Questo può aiutare a capire regole generali per rendere il sistema funzionale.

4.1 Reti metaboliche

Immaginiamo che si vuole curare una certa patologia che è dovuta al comprotamento abnorme di una certa proteina, troviamo una molecola che blocca quella proteina. Si hanno una serie di effetti collaterali. Se un giorno riuscissimo realmente a simulare il metabolismo non avremo bisogno di fare i test clinici. L'unico modo per valutare il tipo di effetti collaterali che si può aspettare, o lo si da ai volontari, o si utilizzano gli animali, il più possibile rappresentativi del caso. Questo tipo di studi che è agli albori sarebbe il sogno di tutti, quello di avere l'intero metabolismo modellizzato. Abbiamo sia reazioni cataboliche che reazioni anaboliche. E ogni freccia è una reazione chimica, ciascuna di queste

reazione è catalizzata dalla presenza di specirfiche enzime. Questa reazione è catalizzata da enzimi. I nodi sono molecole (cataboliti) e le frecce sono gli enzimi.

Questo è un sistema dinamico ma che deve possedere uno stato stazionario. Se facciamo l'ipotesi di essere nello stato stazionario, che ci sia un bilanciamento delle masse, possiamo capire se si riesce a simulare dal punto di vista matematico le reti metaboliche. Se abbiamo dei dati che la nostra rete non riproduce vuol dire che ci manca un pezzo In genere si nutilizza l'analisi del flusso bilanciato.

Si usa la stechiometria delle reazioni, descriviamo il sistema sottoforma di vettori e matrici, supponiamo di avere 4 meaboliti (A, B, C,D), e delle rezazioni:

$$2A + B \to C \qquad \begin{pmatrix} -2 & -1 & 1 & 0 \end{pmatrix}$$

$$C \to D \qquad \begin{pmatrix} 0 & 0 & -1 & 1 \end{pmatrix}$$

I vettori a destra mi rappresentano le reazioni chimiche, in cui i segni negativi sono i reagenti (che si distruggono nella reazione) i segni positivi i prodotti, che vengono generati dalla reazione. Posso scrivere una matrice composta da questi vettori (per righe metabolita e per colonna le possibili reazioni). Se conosciamo la rete metabolica e la stechiometria, si può sempre rappresentare il metabolismo sottoforme di matrice. Possiamo schematizzare i flussi entranti e uscenti dalla cellula come dei vettori nella stessa rappresentazione. Voglio fare in modo che il prodotto della matrice per i flussi sia pari alla variazione del tempo, e voglio che questa variazione sia nulla. Quindi il prodotto righe per colonne della matrice, con i flussi entranti - uscenti deve essere nulla.

Le equaizoni che abbiamo sno quelle che fanno in modo che la concentrazione dei metaboliti rimanga costante, perché siamo allo stato stazioario.

Dove si prendono le informaizone sulla topologia delle reti? Ci sono delle banche dati che ci dicono ciascun enzima quale reazione catalizza. Ciascun enzima è identificato da 4 numeri, il primo numero dice la reazione che catalizza ...

Se abbiamo il numero EC, sappiamo che reazione catalizza l'enzima, se la nostro enzima non è presente, si possono sempre fare una ricerca in banca dati per omologia. Esistono delle banche dati vere e proprie per il metabolismo.

La matrice stechiometrica può essere molto grande, questo non complica particolarmente. Questa reazione quanto consuma di A, quanto di B, quanto produce. Una volta che abbiamo imposto che si tratta di stato stazionario. Dobbiamo risolvere l'equazione omogenea:

$$S \cdot V = 0$$

Dove S è la matrice stechiometrica, e V è il vettore dei flussi. In genere il numero di reazioni è minore del numero di metaboliti, per cui questo sistema è stazionario solo se i flussi soddisfano determinati vincoli.

Abbiamo un intero spazzio delle soluzioni che non sono uniche, ovviamente si possono sempre trovare, qualunque punto in questa regione è una combinazione lineare di quelli che chiamiamo modi fondamentali. Però le risorse dell'universo sono finite, quanto glucosio può entrare in unca cellula? Quanto ossigreno può entrare o uscire nella cellula per unità di tempo? Quindi avremo altri vincoli e altri liminti. Un organimso non ha tutti i flussi dell'universo avrà una soluzione di stato stazionario. Cosa cerca di ottimizzare l'organismo? Minimizzare il consumo? Ottimizzare la crescita?

Una volta che abbiamo lo spazzi delle soluzioni, si possono imporre delle condizioni, se voglio massimizzare la prpoduzione di ATP, che flusso di entrata mi serve? Se finisce all'interno della regione permessa, quello è lo stato che voglio individuare. Abbiamo la rete metabolica che possiamo trasforare nella matrice stechiometrica, Abbiamoil cono, e troviamo il flusso ottimale a secondo della funzioe che riteniamo la cellula voglia massimizzare.

Per testare se la rete metabolica che abbiamo costruito è una rete reale dell'organismo, si procede ad una analisi in silico: si eliminano progressivamente le connessioni e si vede cosa succede ai flussi stazionari, siccome uno dei flussi è vout, questa possiamo misurarla, Possiamo quindi inibire alcune reazioni nel metabolismo, e vedere come si modificano i flussi uscenti, confrontando esperimenti con dati. Questo processo è fatto con il knock out, cioè eliminando sperimentalmente gli emzimi che catalizzano la reazione. Abbiamo un sistema wild-type, e il nostro mutante, ricalcolando i flussi, possono succedere varie cose, quindi, il sistema riesce ad avere, se la nostra ottimizzazione finisce da un altra parte in quelle condizioni non è possibile avere un insieme di flussi che faccia in modo che il sistema risponda nello stato stazionario. Ovviamente esistono dei siti web che permettono di fare analisi online, gratis, in cui gli si da la matrice stechiometrica e lui ci dice cosa volgiamo ottimizzare. Occorre avere uno schema metabolico. Dopo di che se abbiamo uno schema metabolico possiamo fare l'assunsione dello stato stazionario l'assunsione di metaboliti all'interno della cellula rimane costante.

Esiste la pressione osmotica, non si può aumentare eccessivamente all'interno della cellula altrimenti l'acqua esplode. Ciascuna delle nostre cellula ha necessità, abbiamo bisogno una grande quantità di glucosio, nella cellula vogliamo fare in modo abbastanza glucosio, se tenessimo nella cellula tutto il glucosio che servirebbe, quello che facciamo e polimerizzare la molecola di glicogeno, per ridurre la pressione osmotica. Le piante fanno la stessa cosa con l'amido. La concentrazione deve rimanere costante, salvo caso come quello del glicogeno in generale è abbastanza affidabile, che la concentrazione delle molecole rimanga costante non è un assunsiuone folle, sperimentalmente bisogna aspettare che il sistema raggiunga lo stato stazionario.

4.1.1 Inferire il network

In un mondo ideale vorremo scrivere tutte le equazioni differenziali che descrivono il sistema.

Il problema è che non abbiamo tutti i dati necessari in queste equazioni differenziali, hanno un numero di incognite troppo alto. Generalmente si riescono a risolvere per sottosistemi del metabolismo cellulare.

Cosa ci serve per simulare questo sistema? La rete, una serie di reazioni, e poi dei vincoli (il bilanciamento di massa e qualcosa sui flussi). Possiamo simulare la rete per inferire delle proprietà del sistema, e farne una verifica sperimentale. Quando guardiamo una rete ci sono dei flussi: se A viene trasformato in B c'è un flusso da A a B. I flussi possono essere accoppiati (Φ_A determina Φ_B e viceversa) oppure essere disaccoppiati (parti indipendenti del sistema).

Il flusso può essere direzionale (Φ_A determina Φ_B ma non viceversa). Poi abiamo il vilanciamento della massa. Si possono rappresentare le reti con le matrici stechiometrica. Si mettono tutti i nodi (metaboliti) nelle colonne, e le reazioni sulle righe. Se produce un metabolita e produce quel metabolita metto 1, -1 se è consumato, 0 se non succede nulla.

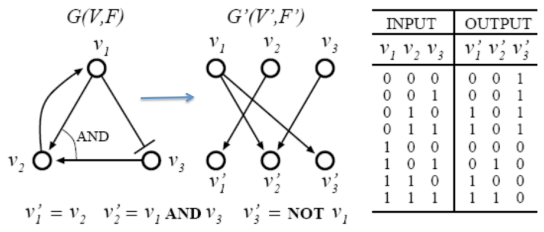
Ci aspettiamo che moltiplicando la patrice stechiometrica per i vari flussi devo ottenere zero (situazione stazionaria). Questo diventa un sistema che ha m equazioni e n incognite, se avessimo m ¿ n avremmo risolto il problema. Generalmente non è questo il caso. Quindi occorre usare dei vincoli per limitare lo spazio delle soluzioni. I flussi devono avere dei limiti, quindi di fatto conosciamo limiti entro cui la soluzione deve trovarsi. Lo scopo della cellula è quello di aumentare la sua massa? voglio allora trovare la soluzione che massimizza la massa nel mio sottospazio di soluzioni accettabili. Se questo è un sistema biologico immagino che la situazione sia tale che cambiando di poco i flussi la situaizone non deve cambiare tanto.

Possiamo utilizzare degli esperimenti per capire se il sistema sta funzionando o meno, ci sono sane persone che hanno eliminato 1 ad 1 tutti i geni guardando se l'organismo sopravviveva o meno. Se nel sistema tolgo un gene mi aspetto che lo spazio delle soluzioni vuoto corrisponda alla morte dell'organismo. Esistono anche particolari coppie di geni che se si elimina un gene e l'organismo sopravvive, elimino l'altro gene l'organimso sopravvive, elimino entrambi i geni e l'organismo muore. Tutte queste informazioni possono essere utilizzate per controllare se effettivamente la rete mi da la reale soluzione.

4.2 Reti Booleane

Le reti booleane consistono nell'approssimare il sistema in modo discreto. Invece di tentare di modellare la variazione dell'espressione dei geni si può fare la seguente approssimazione, un gene p o è attivo o non è attivo.

Stiamo discretizzando il sistema. Un gene può essere identificato con un bit, 1 o 0. Avremo un insieme di nodi (geni) e un insieme di stati (0 o 1) a seconda se i geni sono o meno attivi. Poi abbiamo una serie di funziuoni booleane tra i geni. Si può trasformare l'intera rete in un diagramma booleano. Si ricava una matrice booleana in cui si ricava dato lo stato dei tre nodi in input, qual è lo stato dei nodi in uscita. In altre parole abbiamo una tabellina booleana (Figura 4.2).



INPUT			OUTPUT		
v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figura 4.2: Trasformare una rete in un diagramma booleano.

Come si fa a capire come funzionano queste reti? Dati N nodi, posso avere 2^n stati differenti del sistema (troppi per valutarli tutti). Tuttavia posso partire da un valore casuale e vedere quando si arriva a convergenza: Assegno un valore casuale ad un gene, e vedo l'effetto di questa configurazione in uscita, adesso evolvo ad una situazione successiva. Siccome il numero di stati è finito, una qualunque traiettoria si sceglie, in un certo numero di passi finirà in uno stato che ha già visto. E costruiremo dei cicli. Questi cicli si chiamano attrattori. Si può immaginare che il ciclo attrattore sia un ciclo particolarmente rilevane, perché viene percorso dal sistema più e più volte. Siccome questa rete booleana è deterministica ad un certo punto finiremo per visiare uno stato già visitato, da quel momento ritroveremo quello stato dopo un certo numero di passi.

4.3 Entropia

L'entropia di Shannon in teoria dell'informazione mi da una misura sull'incertezza media della variabile:

$$H(x) = -\sum p(x) \ln p(x)$$

È il limite al di sotto del quale non si può scendere.

Si può definire anche l'entropia di una coppia di variabili:

$$H(X,Y) = -\sum p(x,y) \ln p(x,y)$$

E quindi anche l'entropia condizionale:

$$p(X,Y) = p(X)p(X|Y) \qquad \ln p(X,Y) = \ln p(X) + \ln p(X|Y)$$

$$H(x,y) = H(x) + H(y|x)$$

L'incertezza della coppia è l'incertezza di una variabile, sommata all'incertezza dell'altra variabile data la prima.

La mutua informazione è definita come:

$$M(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \qquad M(x,y) = H(x) + H(y) - H(x,y)$$

Ci da informazione sulla distribuzione corretta.

Si può usare la mutua informazione per costruire la rete booleana. Se vediamo che la conoscenza di una terza variabile mi diminuisce l'entropia relativa ad una coppia di variabili posso sostenere che c'è una connessione.

L'informazione se conosco i primi due nodi è minore o ugale se conoscessi il valore di un terzo nodo? Il funzionamento in dettaglio è riportato in Figura **??**.

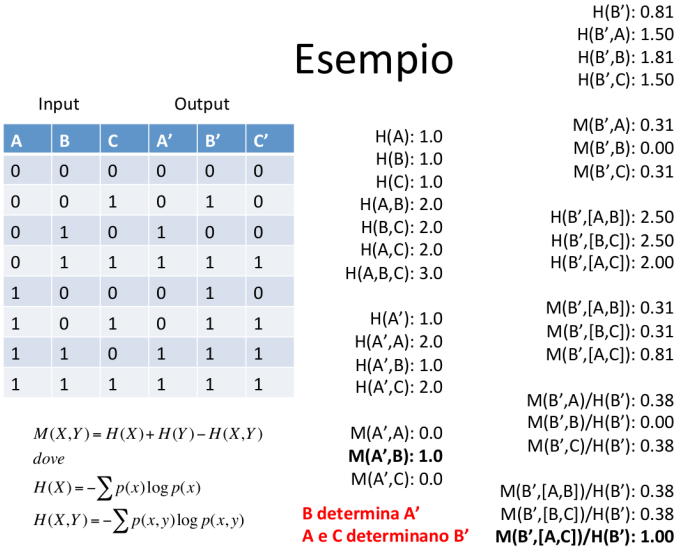


Figura 4.3: Il calcolo dell'entropia può essere usato per inferire le reti.

In queste reti è ingorato sia l'effetto di geni intermedi, ne c'è alcuna considerazione del rumore. Per migliorarle esistono le reti booleane probabilistiche.

4.4 Reti Bayesiane

Le reti Bayesiane sono probabilistiche, può gestire variazioni di input e output e può essere usato in modo molto potente.

Immaginiamo di avere dei dati D e un modello M. Raccogliamo la serie di dati e ci chiediamo se i dati fittano con un modello. Quella è la probabilità dei dati dato il modello $p(D|M)$. Possiamo chiederci qual è la probabilità del modello visti di dati. Il teorema di Bayes ci dice che:

$$p(D,M) = P(M|D)p(D) = p(D|M)p(D)$$

Dato un insieme di dati e un modello midice la probabilità che quell'modello sia corretto. La probabilità del modello entra al numeratore. Questa è detta prior, è una probabilità a priori che assegno al modello. E $p(D|M)$ è la somiglianza tra dati sperimentali e modello. Qual è il vantaggio di questa cosa? Il vantaggio è che si possono suddividere un complesso calcolo delle probabilità in più pezzi.

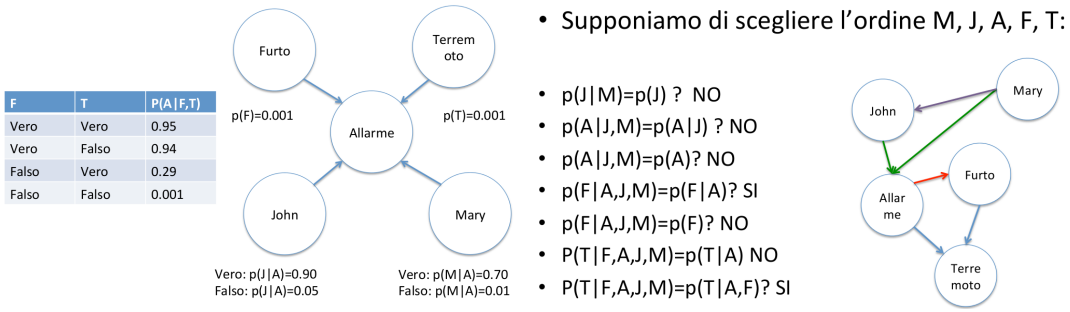
La forza delle reti Bayesiane consiste nel fatto che la probabilità di un nodo figlio dipende solo da quella dei genitori. La probabilità che un gene sia attivo dipende solo dal diretto genitore e quindi possiamo schematizzarla come un network. Possiamo calcolare a pezzi la probabilità, si possono calcolare gli stati intermedi. Il grafo che si ottiene deve essere DAG (*Directed acyclic graph*) senza cicli. Due dati sono *condizionalmente* indipendenti se sono indipendenti dato lo stato di una altra variabile.

Se si vuole testare la qualità dei modelli, o vogliamo selezionare il modello che meglio fitta i dati devo avere una buona distribuzione di probabilità a priori. Altrimenti assegno uniformi le probabilità e ediamo quali di queste meglio rappresenta i miei dati.

Ad esempio se la variabile A assume i tre possibili valori a_1, a_2 e a_3 la probabilità che la variabile B sia pari a b dipende dal valore di a con una certa funzione. Posso cercare di vedere quale modello (funzione) ottimizza il set di dati che conosco. Le variabili possono essere indipendenti o dipendenti a seconda di cosa sto osservando. Ho un modello che dice che date due proteine le carattersitiche importanti per identificare se le i siti interagiscono. Interagiscono in una regione che deve essere conservata evolutivamente e idrofobica (se non non si appiccicano). Qual è la probabilità che quell'aminoacido sia conservato se è un sito di iterazione, la stessa cosa si può fare con l'idrofobicità si possono ricavare le probabilità e si può capire qual è l'informazione di ciascuno di questi valori.

Questo modo di rappresentare le reti consente di fattorizzare la probabilità in modo che una volta noto un valore non dobbiamo più preoccuparci dei genitori. Quanto è probabile che il modello sia corretto? La probabilità dati dato il mdoello è il prodotto della probabilità di cascun dato, dato il modello. Questo valore può essere calcolato, ed è la massima "verosimiglianza". Abbiamo una rete, calcoliamo la verosimilgianza, si ottimizza con vari metodi tipo moltecarlo, come abbiamo visto per costruire gli alberi ottimali, fin tanto che non si ottiene una verosimiglianza massima. Abbiamo una rete, descritta in termini probabilistici, utilizzando il prodotto delle probabilità che ciascun dato fitti il modello e possiamo accettare il nuovo modello se la verosimiglianza è maggiore con metodi di tipo montecarlo.

Il metodo di Bayes può essere usato per inferire, la rete. Vedere Figura 4.4:



(a) prior (b) inferenza

Figura 4.4: Schema di inferenza usando il modello Bayesiano. Si possono cercare altri modelli (sequenze) che che mi massimizzano le probabilità, fino a trovare la rete che mi descrive il corretto funzionamento del sistema.

Costruiamo la rete controlliamo la mutua informazione controlliamo se ci sono variabili diendenti o indipendenti, in questo. La potenza dell'utilizzare il metodo bayesiamo è in due aspetti, fattorizzare la probabilità guardando solo ai genitori di un evento e non tutta la rete. L'altra ragine è che è probabilistica.

Abbiamo un certo numero di variabili, prendiamo una variabile alla volta, e cerchiamo quali dei genitori spiegano completamente la probailità, ad esempio un suo vicino chiama Jon per dirgli se è scattato l'allarme. Possiamo partire dalla rete che conosciamo e calcolare svariati parametri, simulare il comportamento della rete e vedere se fitta i dati sperimentali che abbiamo, oppure se occorre ridisegnare la rete.

Come si costruisce una rete? Possiamo costruire la rete booleiana, in questo caso un nodo è acceso o spento, simulare il comportamento della rete in funzione della tabella che alla fine necessariamente formerà dei cicli. La rete booleana la ricostruisco usando il concetto di mutua informazione, quando 1 da informazioni su un terzo o quarto. Questo metodo è abbastanza veloce ma ha dei difetti, il concetto di dire On-Off senza tener conto di nessun rumore. Il modo più sensato al momento per simulare il comportamento di una rete biologico è quello di utlizzare un metodo probabilistico, si può velocemetne calcolare la verosimiglianza di un certo modlelo di rete con i dati che abbiamo, possiamo invertire le connessioni, una volta che abbiamo la funzione da ottimizzare possiamo utilizzare il metodi di ottimizzazione della verosimiglianza.

Il vantaggio delle reti bayesiane è proprio quello che ricalcolare la verosimiglianza è molto più semplice modificando la rete, poiché cambia solo localmente.

Capitolo 5

Image Processing

L'immage processing si articola in 4 fasi:

- Digitalizzazione
- Compressione
- Enhancement
- Segmentation (Edge Detection)

Per digitalizzare l'immagine si usa una griglia discreta nel piano. E l'immagine sarà campionata nei punti di intervallo:

$$x = j\Delta x \qquad y = k\Delta y$$

Con Δx e Δy intervalli di campionamento.

Il campionamento ideale si rappresenta come una funzione di Dirac:

$$s(x,y) = \sum_{j,k=1}^{M,N} \delta(x - j\Delta x)\delta(y - k\Delta y)$$

L'immagine digitalizzata si ottiene come prodotto dell'immagine originale con la funzione di campionamento.

La funzione di campionamento è una funzione periodica con periodo x, y, e quindi può essere espansa in una serie di Fourier.

Il *pixel* è la minima unità di campionamento.

Si può ottenere un istogramma di un immagine, valutando quanti pixel che hanno intensità compresa in un intervallo di suddivisione ci sono nell'immagine.

Possiamo operare trasformazioni su un immagine attraverso un Kernel g (matrice per la convoluzione:

$$g(m,n) = w(k,l) \otimes f(m,n) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w(k,l) f(m-k, n-l)$$

L'immagine può essere distorta, spesso quello che dobbiamo fare è quello di trovare una funzione tale che applicando questa funzione alla immagine eliminiamo la distorzione ed è invertibile.

Come si fa a trovare la funzione da usare per eliminare la distorsione? Si può fare una rinormalizzazione, se abiamo ad esempio un grande background di pixel che non danno informazione. Se siamo interessati al range tra f_1 e f_2 cremiamo l'immagine così:

$$e = \begin{cases} f & f_1 < f < f_2 \\ 0 & \text{altrimenti} \end{cases} \qquad g = \frac{e - f1}{f_2 - f_1} \cdot f_{max}$$

Dove e è l'immagine intermedia, g l'immagine finale. (Figura 5.1).

Non è detto che sappiamo a priori quale range di intensità ci interessa. Quello che si può fare è di ridistribuire l'instensità su tutti i pixel. In modo da ridistribuire uniformemente le intensità. Questo si fa usando l'istogramma comulativo.

Come si può fare a ridurre il rumore di fondo. Ad esempio si possono fare esperimenti in cui si sequenzia il DNA per poter estrarre informazione in modo corretto bisogna cercare di eliminare il rumore. In generale quello che si fa nel caso più semplice è che a partire dalla griglia si sceglie un sottoinsieme della griglia si attribuisce al pixel centrale il valore medio (Filtro mediano).

L'altro aspetto di un immagine che vogliamo ottenere è quello di identificare specifiche regioni, cioè immagini che sono dentro altre immagini. Ad esempio siamo interessati al riconoscimento dei bordi dell'immagine. Intuitivamente

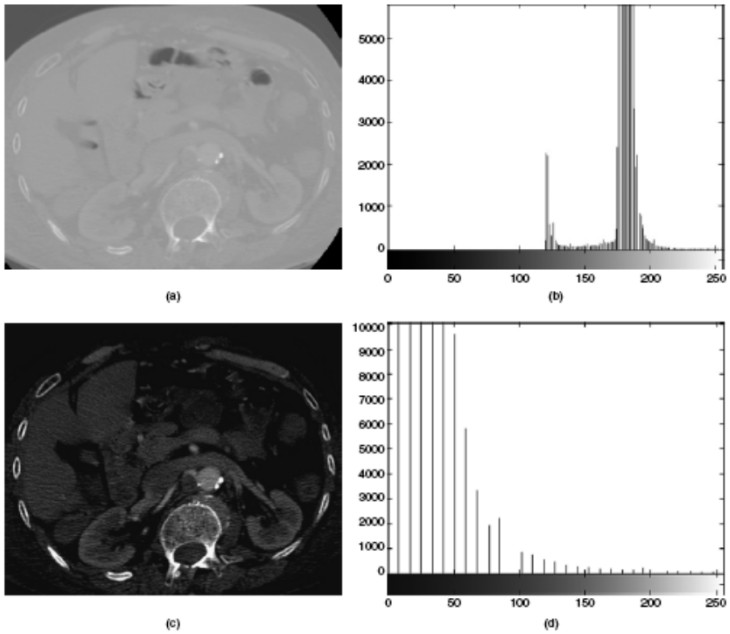


Figura 5.1: Esempio di miglioramento dell'immagine usando lo scaling.

si ha una certa intensità e poi c'è un punto in cui l'intensità cambia, e quello è il punto che vogliamo verificare. Al contorno, la frequenza dell'onda che corrispondono al contorno sono in genere alte frequenze. Se l'approssimiamo con la trasformata di fourier in quella regione le frequenze sono più alte. Possiamo usare le derivate o un kernel apporpirato.

Si possono usare i kernel. Un kernel è la matrice che si può usare per trasformare i dati:

$$w_{H1} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} \qquad w_{V1} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

Un altra cosa che spesso deve essere fatta è quella di sottrarre un background. Ci sarà un background di luce che è diverso asseconda dei punti che stiamo considerando. Posso usare la convoluzione tra fuzione rumore e immagine, fin quando non si ottiene una cosa che si desidera. Se un numero grande di immagine dello stesso tipo, e le immagini sono affette da un rumore casuale a media nulla, possiamo mediare su tutte le immagini in modo da ripurirle.

Possiamo voler migliorare ulteriormente il contrasto dell'immagine. So può fare un istogramma. C'è bisogno di qualcuno che osservi l'immagine. Selezzioniamo due punti, uno dell'immagine e uno del backgraund, e espandermi attorno a questi punti e calcolare l'istogramma su questi pixel per eliminare la parte di istogramma relativo al background.

5.1 Region growing

Possiamo essere interessati a gruppi di pixel che hanno intensità simile in modo da riconoscere strutture all'interno dell'immagine. Si inizia da un pixel o da un gruppo di pixel (detti seed). Ogni pixel nuovo viene esaminato e aggiunto alla regione se sufficientemente simile al seed.

Si può anche calcolare il gradiente dell'immagine, e quando la derivata prima raggiunge un massimo, quello può essere il punto in cui finisce l'immagine e inizia il background. E voglio trovare quali sono i gruppi di pixel che hanno intensità simile. Scegliamo un pixel centrale, o si fa automaticamente. E poi si comincia a crescere guardando i pixel limitrofi se è sufficientemente simmile lo aggiungiamo all'immagine o lo lasciamo fuori. Questi risultati psssono essere anche ottenuti con metodi di apprendimento automatico, e chiedere al sistema automatico di individuare immagine e background. Queste cose si fanno e sono poco utilizzate nella letteratura medica perché nessun metodo dell'apprendimento automatico è molto accuramento. I

n questo modo posso identificare i limiti di un oggetto. Ci sono dei pesci che i predatori riescono a vedere l'ombra delle prede proiettate sul fondo. Il pesce preda luminesce in modo da eliminare l'ombra. Questa proteina si chiama green flourescence protein, ed assorbe ad una certa lunghezza d'onda, si conosce la struttura della proteina, abbiamo ricavato un sacco di colori modificando gli aminoacidi intorno a quella zona. Queste proteine le possiamo attaccare alla fine del gene della proteina che ci interessa. Possiamo quindi tentare di sovrapporre le due immagini, per vedere se i due oggetti biologici che abbiamo marcato con i due colori hanno una certa correlazione nel posizionarsi.

Date due immaginisi può ottimizzare il coefficiente di correlazione, qual è la migliore traslazione e rotazione che permette di aumentare la sovrapposizione tra le due immagini? Si può ottimizzare rispetto a traslazioni e rotazioni, si prova con rotazione e traslazioni scala, deformazioni fino ad ottenere il massimo della correlazione.

Per migliorare il confronto si può anche utilizzare la mutua informazione: che informazione ho sulla seconda immagine se conosco la prima. Possiamo ottimizzare la mutua informazione, possiamo volere il caso in cui la conoscenza di x mi specifichi del tutto y.

$$h(x) = -\int p(x) \ln p(x) dx \qquad h(x,y) = -\iint p(x,y) \ln p(x,y) dx dy \qquad I(x,y) = h(x) + h(y) - h(x,y)$$

Un altra particolare implicazione è cercare se c'è un metodo per tracciare i percorsi fatti dalle cellule attraverso tane immagini. È fattibile come cosa, bisogna assumere che la maggiorparte delle cellulce non si sono spostate e ottimizzare il minimo numero di cellule che si muovono, e tra quelle spostate si può fare un tracking del loro movimento.

Un altra tecnica molto interessante è la ricostruzione tridimensionale.

Capitolo 6

Metodi sperimentali

I metodi sperimentali che consentono di risolvere la struttura tridimensionale delle proteine sono principalmente i seguenti:

- Microscopia elettronica
- Cristallografia a raggi X
- Risonanza magnetica nucleare (NMR)

6.1 Microscopia elettronica

La Microscopia Elettronica permette di arrivare ad una risoluzione di circa 0.05 nm, ben superiore al limite ottico di diffrazione di circa 250 nm (lunghezza d'onda del visibile).

Il campione deve essere preparato, abbiamo due metodi principali il SEM (Scanning Eletronic Microscopy) in cui il campione è coperto da un metallo che riflette gli elettroni, e impedisce che il campione si carichi (si osserva solo la superficie del campione, e il TEM (Trasmission Elettronic Microscopy) in cui gli elettroni passano attraverso il campione e si ottengono immagini bidimensionali (il campione deve essere sottile).

Il campione deve essere raffreddato con azoto liquido. Per aumentare il contrasto dell'immagine sono usate tecniche come lo Staining Negativo, in cui si sostituisce il background con del colorante per aumentare il contrasto dell'immagine (negativo perché agisce sull'ambiente e non sul campione). Questo processo può distorcere però l'immagine che si genera e introdurvi artefatti. Per evitare questo si può usare la tecnica di raffreddamento del campione, in cui il campione viene congelato[1].

Il processo di misura genera tante immagini che sono proiezioni del nostro campione viste da varie angolazioni. Dobbiamo quindi disporre di un algoritmo utile per estrarre da queste proiezioni un modello tridimensionale. La logica che segue l'algoritmo è di tipo iterativo. Prima vengono prese le proiezioni e a partire da questo si sviluppa un modello preliminare di vista 3D, poi usando questo modello si raffina nuovamente il processo, pescando nuove particelle, ecc.

Poiché le immagini del campione inclinato si disporranno su dei coni concentrici (Figura 6.1) siamo in grado, selezionata la particella non inclinata di stabilire per ogni immagine la sua coordinata θ e φ da cui il campione è stato "fotografato".

Perché la ricostruzione di immagini funzioni correttamente è importante sapere esattamente il punto in cui è stata fatta la proiezione. Quindi c'è bisogno di un algoritmo che rifinisca le posizioni dell'immagine. Questo viene fatto con un algoritmo ricorsivo. Per ciascuna immagine viene calcolato il centro di massa dell'intensità e poi mediato con tutte le altre particelle. A questo punto si sposta l'immagine in modo che ciascun centro di massa sia il più vicino alla regione del cono. A questo punto si riprende come centro della particella quello con cui intercetta il cono di proiezione e si rimedia su tutte le particelle per ottenere una nuovo centro di massa, e si procede iterativamente fin quando ciascuna particella non viene più spostata.

A questo punto le immagini vengono classificate in gruppi in cui si vedono sulla stessa orientazione.

Dopo la classificazione si procede a fare un modello tridimensionale dell'immagini dai dati (usando la trasformata inversa di Rodon), questo modello viene quindi usato nuovamente per catalogare meglio i dati, e rifare nuovi modelli fin quando non si arriva a convergenza.

Per ricavare come effettivamente la struttura della proteina si dispone in questa ricostruzione tridimensionale si può usare una simulazione di dinamica molecolare in cui si aggiunge al potenziale dei termini elastici che sono minimi se è massima la sovrapposizione tra modello e figura sperimentale. In questo modo si è risolta la struttura tridimensionale della proteina.

[1] Anche questo processo ha i suoi svantaggi: il campione in questo stato non può essere esposto più di una volta al fascio di elettroni altrimenti viene distrutto.

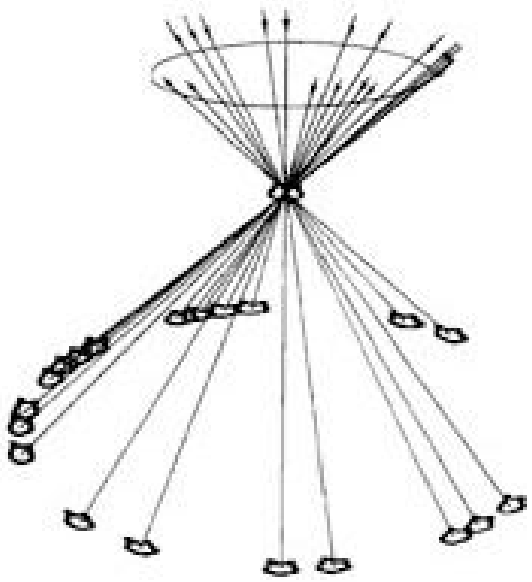


Figura 6.1: Esempio di come sono generate le immagini proiettate dal campione.

6.2 Cristallografia a raggi X

La cristallografia a raggi X è una tecnica molto utilizzata: La radiazione X ha lunghezza d'onda dell'ordine del Armstrong, permettendo quindi rispetto alla luce visibile sia maggiore penetrazione ($E = h\nu = \frac{hc}{\lambda}$) sia maggiore risoluzione (il limite di diffrazione è dell'ordine dell'armstrong).

È però molto difficile costruire lenti per i raggi X e lo scattering da un singolo atomo è debole, per cui occorre che il campione sia in un cristallo, con molte molecole nello stesso orientamento.

Si registra la radiazione diffratta per calcolare l'immagine, purtroppo il pattern di diffrazione mi mostra solo l'intensità non la fase dell'onda, impedendomi a priori una ricostruzione completa dell'immagine.

Per cristallizzare le proteine sono stati inventati numerosi metodi: si mette la proteina in una goccia di soluzione e si sopra un sottile strato di copertura, con acqua sotto, in modo che l'acqua evapori dalla goccia, con l'aggiunta di un precipitante.

Aggiungendo un precipitante e aumentando la concentrazione della proteina si arriva nella fase in cui la proteina inizia a precipitare (sovrasaturazione). Ma se la proteina inizia a cristallizzare, la sua concentrazione diminuisce. Tra i precipitanti più usati c'è l'ammonio solfato e NaCL.

La presenza del cristallo può essere verificata sfruttando la birifrangenza. Se all'interno del campione si formano microcristalli bisogna cercare di vedere se effettivamente sono cristalli di proteina o solo sale.

Quando si inviano i raggi X al campione la diffrazione fa interferenza costruttiva agli angoli di Brag:

$$n\lambda = 2d \sin\theta$$

Dove d è la distanza tra due piani cristallini. Ciascun punto nel paterne di diffrazione ci dice quanto sono concentrati gli oggetti sui piani corrispondenti.

Quando il raggio incidente incontra un set di piani che diffrangono, il suo vettore reciproco finisce nella sfera di Ewald.

L'onda incidente possiamo schematizzarla come un onda piana:

$$Ae^{i\vec{k}\cdot\vec{r}}$$

Se in $\vec{r}$ c'è un campione di densità $f(r)$ allora l'onda diffratta sarà:

$$A \cdot Sf(r)r^{i\vec{k}\cdot\vec{r}}dV$$

Il fotone percepito nel rilevatore avrà una fase variata nel seguente modo:

$$e^{i\vec{k}_{out}(\vec{r}-\vec{r}_{screen})}$$

L'informazione sul campo elettrico è data da:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl}|F_{hkl}|e^{-2\pi i(hx+ky+lz)+i\varphi_{hkl}}$$

Da cui Facendo la trasformata inversa ricaveremmo l'immagine nello spazio reale. Tuttavia misuriamo solo l'intensità:

$$I \propto |F_{hkl}|^2$$

Per risolvere questo sistema si possono usare vari modi. Un modo è quello di aggiungere all'intensità una fase nota che ci aspetteremmo da qualcosa di simile alla molecola. Ad esempio come mostrato in questo sistema 6.2

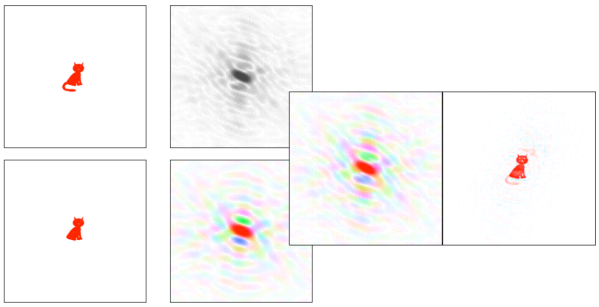


Figura 6.2: Del gatto con la coda abbiamo solo l'intensità della trasformata di Fourier, del gatto senza coda abbiamo anche la fase. Aggiungiamo la fase del gatto senza coda all'intensità del gatto con la coda e facciamo la trasformata inversa ottenendo informazioni anche sulla coda.

Possiamo anche cercare di ricavare la fase ignota "battendo" con un onda di riferimento di ampiezza e fase nota (Rimpiazzo isomorfo), esiste un metodo degli atomi pesanti, in cui vengono aggiunti atomi pesanti (con tanti elettroni) alla proteina, e si creano due pattern di diffrazione. Dalle differenze di questi due pattern è possibile costruire una mappa di Patterson da cui ricavare le fasi.

Abbiamo il metodo del rimpiazzo molecolare. Supponiamo di avere un modello della molecola. Le fasi possono essere calcolate usando il modello e posizionandolo nella cella unitaria. Calcolare quindi il pattern di diffrazione del modello, e usare metodi di correlazione di Patterson per confrontare la mappa calcolata e misurata. A questo punto si raffina il modello in modo da ricavare la mappa.

Una volta che si ottengono le mappe di densità bisogna usare i metodi di raffinamento per poter milgiorare la proteina: si introduce un termine nel potenziale che mi sovrappone la mappa di densità misurata a quella della proteina e quindi si calcola un nuovo modello per la proteina, da usare come base su cui ricavare nuovamente le fasi del campione.

La misura di qualità può essere fatta attraverso l'R-Factor:

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|}$$

O il B factor:

$$B_j = 8\pi^2 \left\langle u^2 \right\rangle$$

Dove $\left\langle u^2 \right\rangle$ è lo scarto quadratico medio dell'atomo dal centro di scattering, dovuto ad agitazione termica. Questo ci da informazioni su quali parti della proteina sono effettivamente più ferme e quali più mobili. L'accuracy è il rapporto tra densità elettronica osservata e attesa.

6.3 Risonanza Magnetica Nucleare

Nella risonanza Magnetica nucleare si osserva il comportamento dei protini, e la proteina è in soluzione.

Non tutti i protoni di una proteina sono equivalenti, la presenza di altri protoni schermano gli effetti del campo magnetico. Questo dipenderà dalla struttura stessa della molecola. Queste variazioni in genere si esprimono per parti su milione rispetto al campo esterno.

Il protone in soluzione ha un certo valore di frequenza di Larmor. Il fatto che ci siano altri protoni intorno modifica leggermente questa frequenza: dipende dal campo magnetico di cui risente questo protone, e il campo magnetico dipende da cosa c'è intorno.

Il protone ha due livelli energetici di spin, la cui differenza energetica cambia a seconda di cosa c'è intorno. Guardando la frequenza di larmor possiamo sapere se il protone è legato ad un carbonio o all'ossigeno.

Abbiamo delle tablline, e sappiamo se l'idrogeno è legato al carbonio in un certo modo ci aspettiamo una data differenza di livelli energetici. Questa differenza sarà una differenza piccola, ma misurabile.

Cominciamo ad inviare al campione delle radio frequenze. Attiviamo il campo magentico e osserviamo la frequenza con cui il protone ritorna lungo l'asse z. Questo succede quando la radiofrequenza che mando è esattamente uguale alla differenza tra i due stati di spin. Possiamo osservare la freuqenza di ciascun atomo attivando il campo magnetico

e osservando la corrente che si genera nel soleniode. Possiamo mandare una radiofrequenza a varie frequenze, e un certo protone genererà una corrente e un altro no.

Ciascun gruppo chimico avrà un caratteristico spettro di assorbimento. Per ciascun aminoacido sappiamo quali valori ci aspettiamo per ciascuna proteina. Abbiamo registrato in che punto veniva assorbita l'energia. Guardando questo patner possiamo identificare i gruppi chimici che hanno assorbito la radiofrequenza.

Immergiamo il campione in un campo magnetico costante B_0. Quindi, infiliamo il campione in un solenoide con asse ortogonale al campo magentico B_0 che genera un piccolo campo B_1. Quando si spegne B_1 lo spin rotea come una trottola attorno a B_0 con la frequenza di Larmor, rilassandosi. Questa progressione induce nel solenoide una piccola corrente che può essere misurata (Figura 6.3):

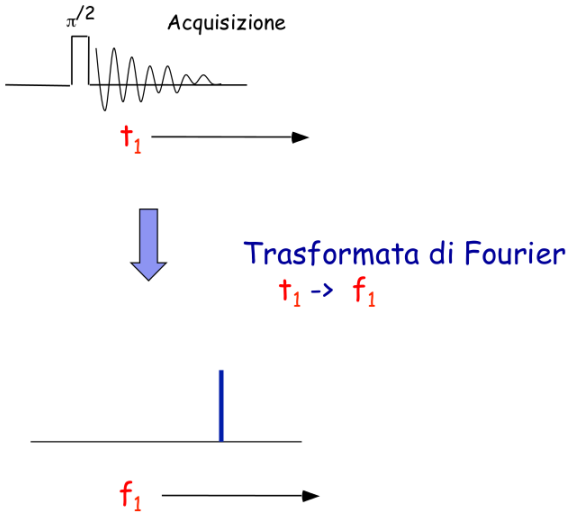


Figura 6.3: Schema di come si ricavano le frequenze di Larmor.

Quello che vorremmo è trovare sono relazioni strutturarli degli aminoacidi. Si può sfruttare anche l'effetto NOE (nuclear overhauser effect).

Irradiamo un protone criamo fino a pareggiare le popolazioni di protoni di quel tipo di spin up e down. A questo punto il sistema non assorbe più la mia radiazione, torna ad assorbirla più gli spin tendono a rilasssarsi. Se i fotoni sono separati l'unico rilassamento possibile è quello che mi ribalta lo spin del protone che avevo eccitato. Se tuttavia due protoni sono ad una distanza piccola allora possono interagire e il mio sistema può decadere in uno stato di minore energia in cui entrambi i protoni vengon invertiti di spin. Questo modifica la popolazione nello stato fondamentale dell'altro protone, facendo diminuire l'intensità nell'assorbimento alla sua frequenza (Figura 6.4).

Poiché l'effetto NOE dipende dalla distanza alla 6 dei protoni possiamo stimare le distanze tra i vari protoni dei vari gruppi chimici (amino acidi).

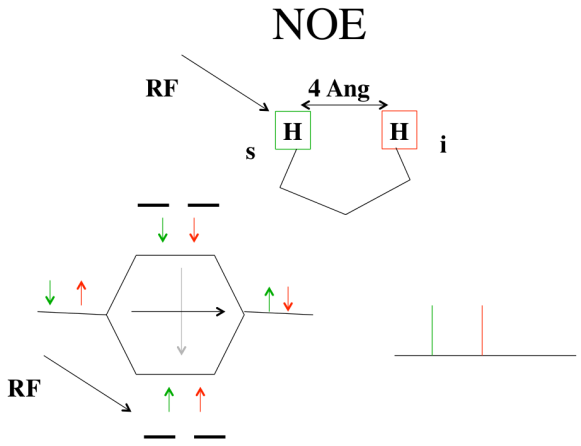


Figura 6.4: Schema dell'interazione dei protoni nel NOE.

- ciascun protone assorbe ad una frequenza asseconda dell'intorno (assegnamo una frequenza a ciasucn protone della proteina)

- Eccitando uno specifico protone (di cui conosco la frequenza di eccitazione) posso vedere come cambia l'intensità degli altri, se l'intensità degli altri cambia esiste un certo range che li separa.
- Genero dei vincoli sulle distanze.
- Ci mettiamo dentro lunghezze, angoli e ricostruiamo la struttura.

La proteina in soluzione si muove, molecole molto grandi che si agitano di meno hanno bisogno di ampi magentici più forti, più è grande il campo magnetio maggiore è la dimensione delle proteine. La proteina deve essere pura. Deve essere abbastanza concentrata.

Se prendiamo una proteina possiamo fare dinamica molecolare assumendo che i vincoli di distanza siano delle molle.

Bisogna trovare un metodo per identificare le strutture / la struttura che soddisfano questi vicnoli. Abbiamo un sistema di distanze non esaustivo, e vogliamo trovare la struttura della proteina che soddisfa questi sistemi.

Un metodo per completare i dati è quello di costruire una matrice con tutti i protoni nelle righe e nelle colonne, per alcuni abbiamod dei dati. Questa matrice è espansa. Abbiamo dei numeri approssimati e neanche completi. Usiamo la diseguaglianza dei triangoli. Possiamo riempire approssimativamente con dei limiti superiori e inferiori, se abbiamo la distanza tra A e B e tra B e C possiamo dare dei limiti a quella tra A e C. A questo punto con questa matrice possiamo minimizzare l'energia soddisfando i vincoli. (Distance Geometry)

Quello che si fa più spesso è che siccome sappiamo calcolare il potenziale della proteina. possiamo aggiungere un termine al protenziale e dire al sistema di esplorare sia dal punto di vista energetico sia dal punto di vista strutturale. Cioè aggiungiamo una pseudo forza non reale tale per cui se due atomi sono ad una distanza di circa qualcosa mettiamo una molla, con delle costanti che possono dipendere dalla variazione di intensità e simuliamo in dinamica molecolare. Spesso si possono scegliere protocolli diversi. Prima si mette un peso molto maggiore ai vincoli NMR, e poi aumentiamo il peso dei parametri energetici. Un problema divena ottimizzare la struttura dati dei vincoli. Il problema è sottodimensionato, quello che si fa alla fine è ottimizzare a partire da condizioni iniziali diverse, poi si possono sovrapporre, se partendo da cose diverse otteniamo strutture molto simili siamo felici.

Abbiamo una rappresentazione delle possibili soluzioni del problema, più sono simili i risultati partendo da condizioni diverse più risultati si vedono. È importante trasmettere l'informazione quanto è buono il risultato. Nel caso dell'NMR si da la radice quadratica media fra gli atomi delle strutture che abbiamo ottenuto. Più bassa è la radice quadratica media delle strutture, più definita è la struttura, più i dati sono sufficienti per arrivare sempre alla stessa sturttura. Se stiamo facendo un esperimento NMR quanti vincoli ci servono.

Quanti vincoli occorrono dipende dalla topoloigia della proteina. Cosa viene fuori se si guarda un articolo NRM, troveremo tante strutture di una proteina. Dopo di che abbiamo dei parametri. Se sono a lunga o a media distanza. Tra un aminoacido e quello successivo sono utili per assegnare a ciascun protone la sua risonanza. Quello che si fa, è che una volta che abbiamo le strutture calcoliamo quanto si accordano coi vincoli. Come si accordano le strutture con i dati è molto importante come dato per capire la bontà della predizione. Quindi spesso si graficano tutte le figure ottenute sovrapponendole (NMR ensamble).

6.3.1 Altri impieghi

Ci sono una serie di altri utilissimi esperimeni che si possono fare con la NMR. Possiamo prendere una proteina destrutturata, e la mettiamo in una soluzione con deuterio anziche acqua. Quello che succede è che i protoni della proteina si scambieranno col deuterio. Facciamo ristrutturatre questa proteina in acqua. I protoni che sono nascosti dall'acqua non scambieranno più invece gli atomi delle regioni più esposti li vedrò per NMR. In questo modo possiamo seguire qual è il pezzo che si struttura prima, quale quello che si struttura dopo, possiamo anche per esempio farla strutturare in deutrerio, aggiungiamo un altra proteina che interagisce con la prima e capire quali sono le regioni di interazione.

Sempre in NMR è possibile prevedere quali composti interagiscono con la proteina Questo era una cosa di dire, la mettiamo nel tubo NMR, poi prendiamo 100 composti e gli aggiungiamo. Se mi cambia lo spettro di uno di questi farmaci vuol dire che si è legato alla proteina. Ci sono un enorme numero di esperimenti che si possono fare con NMR. Si possono identificare i protoni grazie al fatto che i protoni che ci sno intorno schemrano il campo magnetico, e protoni che hanno distanze basse risentono tra di loro e possiamo misurare l'interazione fra i vari protoni e individuare coppie di protoni e entro distanze tipiche.