Il sistema immunitario attraverso la lente dell'inferenza statistica

The men of experiment are like the ant, they only collect and use; the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes a middle course: it gathers its material from the flowers of the garden and of the field, but transforms and digests it by a power of its own.

F. Bacon

Barbara Bravi *Department of Mathematics, Imperial College London, London, United Kingdom*

Il sistema immunitario è capace di mettere in campo risposte estremamente specifiche che, a livello molecolare, si fondano sul riconoscimento degli agenti patogeni esterni. A seguito del recente *boom* nelle tecniche di sequenziamento, è divenuto possibile repertoriare nel dettaglio gli insiemi di proteine coinvolte in tale riconoscimento, producendo così risorse senza precedenti per caratterizzarne quantitativamente le proprietà ed il funzionamento. L'obiettivo di questo articolo è tracciare una panoramica di alcuni approcci di modellizzazione del sistema immunitario che sono basati sui dati di sequenziamento e che uniscono, al potere esplorativo e predittivo dell'apprendimento statistico, l'interpretabilità dei modelli di meccanica statistica. Se da un lato il fine primario di questi approcci è stabilire un quadro di comprensione teorica dei meccanismi di risposta immunitaria a livello microscopico, dall'altro le loro predizioni dimostrano importanti risvolti applicativi nello sviluppo dei vaccini e dell'immunoterapia.

La risposta immunitaria

Il nostro sistema immunitario comprende un insieme di cellule, mediatori biochimici ed organi nel loro complesso adibiti all'individuazione e all'eliminazione di tutti quegli agenti che possono causare malattie.

Per iniziare, vorrei descrivere alcune dinamiche fondamentali tramite cui questa azione viene dispiegata, concentrandomi sui `linfociti T`, ed in particolare sui `linfociti T` di tipo `killer`, e rimandando per una trattazione più dettagliata

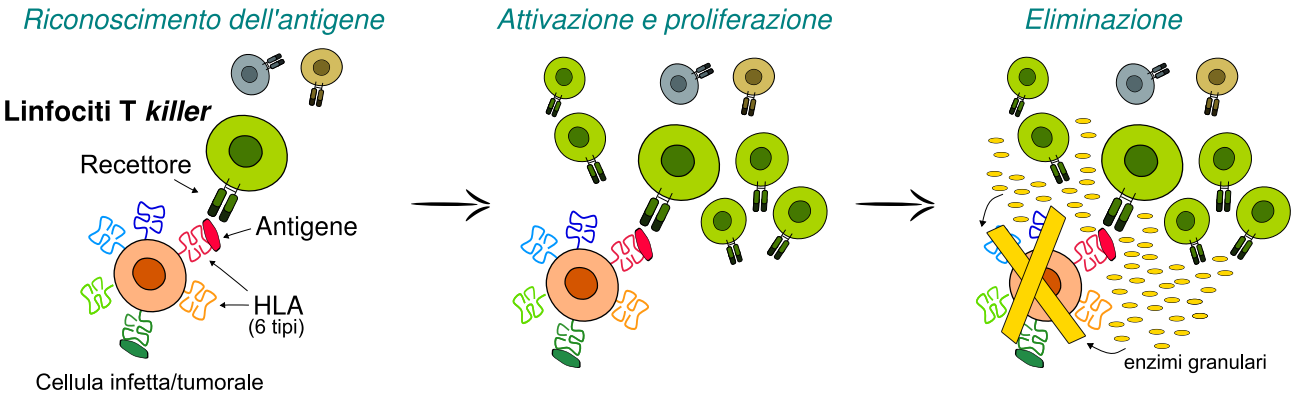

Figura 1: *Meccanismi di dispiegamento della risposta immunitaria da parte dei* `linfociti` *T di tipo killer.*

ed istruttiva sul sistema immunitario a testi come [1].

I `linfociti T`, assieme ai `linfociti B`, sono le cellule centrali del sistema immunitario **adattativo**, cioè la parte del sistema immunitario che, differentemente dalla parte cosiddetta innata, è presente solo nei vertebrati ed è capace di risposte molecolari specifiche. Un meccanismo essenziale sotteso a queste risposte molecolari specifiche è innanzitutto il riconoscimento a livello molecolare di un agente patogeno o di una massa tumorale.

Il riconoscimento specifico

Il riconoscimento specifico di un patogeno da parte dei `linfociti T` *killer* viene iniziato da un legame biochimico fra i `recettori` di membrana dei `linfociti` T e l'`antigene` (Fig. 1). Per `antigene` si intende una porzione di proteina, per esempio delle proteine di un virus che può avere infettato una cellula, oppure delle proteine contenenti alcune delle mutazioni che distinguono una cellula tumorale da una normale; questi frammenti vengono prodotti all'interno delle cellule nel normale processo metabolico di degradazione delle proteine.

Tuttavia gli `antigeni` non sono tutte le possibili porzioni di una proteina, ma sono solo quelle che hanno un'alta affinità di legame con dei complessi proteici delle cellule umane chiamati `complessi HLA di classe I` (dall'inglese *Human Leukocyte Antigen*), specializzati nella cosiddetta **presentazione** dell'`antigene`. La `presentazione dell'antigene` consiste nell'esporlo sulla superficie cellulare affinché possa essere visibile e riconoscibile dai `linfociti T`, i quali si sono evoluti in modo da discriminare, attraverso i propri `recettori` di membrana, gli `antigeni` che provengono da proteine naturali dell'organismo e quelli di origine virale, batterica o tumorale (a cui soli si legano).

Il legame biochimico `antigene-recettore` innesca delle reazioni biochimiche che promuovono l'attivazione e la proliferazione (per divisione cellulare) di quei `linfociti` che possiedono un `recettore` capace del riconoscimento specifico di un certo `antigene` (Fig. 1). In questo modo il riconoscimento dà luogo alla cosiddetta `espansione` di una popolazione di `linfociti` T specifici, la quale andrà a presiedere all'eliminazione delle cellule infette o tumorali attraverso la secrezione di sostanze biochimiche (degli `enzimi granulari`) che ne inducono l'`apoptosi` (si veda ancora Fig. 1).

Esistono varie tipologie di `linfociti` T, i quali svolgono ruoli altamente specifici e fra loro complementari; per esempio, esistono anche i `linfociti` T di tipo *helper*, i quali attraverso il `recettore` riconoscono gli `antigeni` presentati non da cellule infette o maligne bensì da altre cellule del sistema immunitario (come le cellule dendritiche) capaci di intercettare agenti patogeni presenti nell'organismo. Sono denominate *helper* perché producono delle molecole (`citochine`) che stimolano l'attivazione dei `linfociti` T *killer* o degli altri `linfociti` centrali del sistema adattativo, i `linfociti` B (ossia quei `linfociti` dotati di `recettori` che circolano nell'organismo anche in forma solubile, i ben noti `anticorpi`).

Il meccanismo illustrato per i `linfociti` T, nella sua organizzazione generale, è in realtà quello attraverso cui, a tutt'oggi, si cerca di comprendere la risposta immunitaria messa in atto

dall'intero sistema adattativo e che fu proposto da Frank Macfarlane Burnet nel 1957: la cosiddetta teoria della selezione clonale [2]. Il quadro esplicativo di questa teoria si articola in 4 assunzioni ampiamente condivise dalla comunità scientifica: i linfociti T e B attuano un riconoscimento specifico degli antigeni attraverso i propri recettori di membrana, questo riconoscimento seleziona un clone (una popolazione di linfociti con lo stesso recettore) che, ingrandendosi per divisione cellulare, è in grado di rimuovere le cellule infette e permanere a lungo nell'organismo costituendo la memoria immunitaria.

Complessivamente, si può dire che il funzionamento del sistema immunitario coinvolge diverse scale spaziali e temporali e necessita, quindi, anche di diversi livelli di descrizione matematica, come ampiamente discusso per esempio in [3] e nelle referenze lì citate. In breve, si spazia dal livello dell'interazione molecolare, cioè della formazione di legami fra proteine (recettori e antigeni), alle reazioni biochimiche di *signaling* che innescano l'attivazione e proliferazione dei linfociti, alla comunicazione inter-cellulare, mediata da molecole secrete come le citochine e che interviene a coordinare un comportamento collettivo di risposta. Vi si aggiunge poi la scala globale della popolazione umana, che intrattiene una dinamica di interazione con agenti patogeni come i virus, in cui questi ultimi mutano continuamente per evadere la pressione del sistema immunitario umano.

Ciò che è affascinante è che una risposta così specifica e precisamente orchestrata abbia però, come condizione di possibilità, un'enorme diversità di base, che è la diversità, sia a livello intra- sia a livello inter-individuale, dei cosiddetti **repertori immunitari**.

I repertori immunitari e la loro diversità

Per **repertorio immunitario** si intende l'insieme dei recettori di tutti i linfociti per esempio in un individuo, il cui numero è impressionante. Si stima che il repertorio immunitario individuale per i linfociti T consista in $\sim 10^8$ recettori unici, in gran parte **privati** dell'individuo, ovvero non condivisi a livello di popolazione. Questo numero è impressionante specialmente se messo a confronto con il numero totale di geni nell'essere umano ($\sim 2 \cdot 10^4$): non è quindi possibile pensare alla formazione dei recettori puramente secondo il dogma centrale della biologia, per cui un gene ha tutta l'informazione per determinare univocamente una proteina. Se così fosse, servirebbero così tanti geni da non poter essere contenuti nel nucleo della cellula!

La generazione di questa diversità è il risultato di un processo stocastico, noto come ricombinazione V(D)J, che avviene durante lo sviluppo dei linfociti sia T che B [1].

In estrema sintesi, il genoma contiene dei segmenti genici chiamati V, D e J (rispettivamente *Variable*, *Diversity* e *Joining* in inglese), per ciascuno dei quali esistono più varianti (Fig. 2A). Durante la sintesi dei recettori (o, più precisamente, di ciascuna delle due catene costitutive, denotate con α e β per il linfociti T, Fig. 2A), una della varianti per ciascuno di questi segmenti viene scelta a caso e combinata a formare il gene da cui sarà prodotto il recettore. Come ulteriore azione di diversificazione, vengono rimossi (delezioni) o inseriti (inserzioni) dei nucleotidi alle giunzioni fra questi segmenti, ancora una volta in modo aleatorio. Da questo gene ricombinato e modificato in maniera aleatoria viene tradotta la composizione in amminoacidi dei recettori, che è resa così estremamente variabile fra un recettore e l'altro, soprattutto nella regione di giunzione fra segmenti. Questa regione, detta CDR3 (dall'inglese *Complementarity-Determining Region 3*), è proprio la regione del recettore che stabilisce il legame con gli antigeni. Si tratta dunque di un'estrema variabilità giustificata in un'ottica funzionale: la diversità accresce il potenziale di riconoscimento del repertorio, ovvero la sua capacità di intercettare una vasta gamma di possibili antigeni presentati al sistema immunitario.

A seguito della ricombinazione V(D)J, i recettori dei linfociti T generati vengono sottoposti ad un processo di selezione sulla base delle proprietà biochimiche di legame, la quale avviene nel timo (selezione timica). In questo processo, i linfociti T sono messi alla prova a confronto con degli antigeni provenienti dalle proteine naturali dell'organismo, andandoli a scartare ogni qualvolta il legame instaurato con l'antigene è troppo forte (poiché rischiereb-

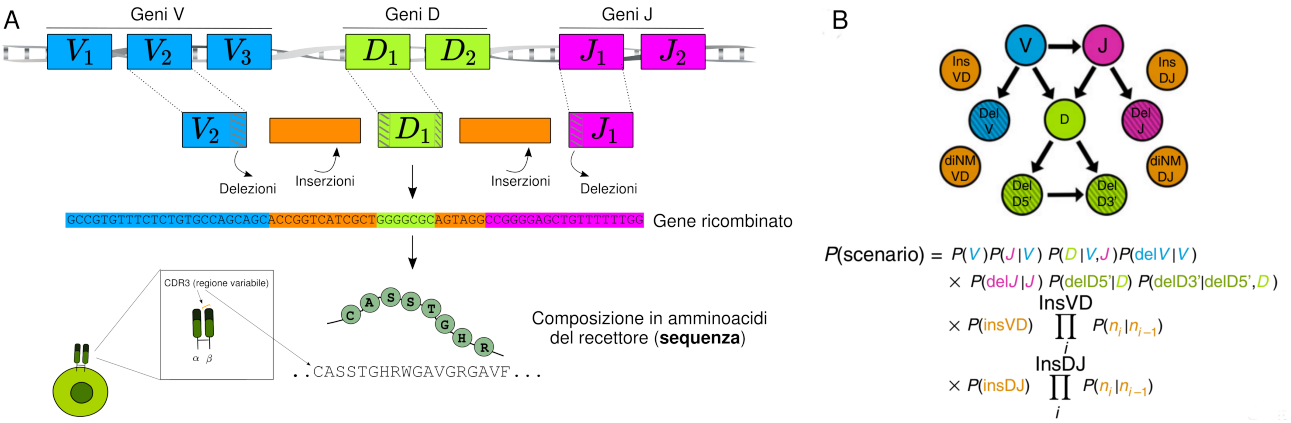

Figura 2: *A: Il processo di ricombinazione V(D)J responsabile della generazione dei* `recettori` *dei* `linfociti`*, qui illustrato per la catena β dei* `linfociti` *T (la catena α è prodotta da una ricombinazione che coinvolge solo segmenti di tipo V e J). B: Codifica dei meccanismi di ricombinazione (scelta dei segmenti VDJ, inserzioni e delezioni) nella formula per la probabilità di generazione utilizzata in [7, 8, 9]; la figura B è presa da [8].*

bero di riconoscere come estranee anche le proteine proprie dell'organismo, provocando così sindromi autoimmuni) o troppo debole (indice del fatto che potrebbero poi non legarsi ad alcun `antigene` patogeno), si veda Fig. 3A.

In maniera speculare, anche gli `antigeni` presentabili al sistema immunitario formano un insieme di una diversità imponente (stimata attorno a 10^{10} sequenze [4]). A loro volta poi, i geni HLA sono altamente polimorfici, con migliaia di varianti distinte esistenti nella popolazione umana. Le cellule di ogni individuo possiedono 6 tipi di proteine HLA potenzialmente diversi (Fig. 1), 2 codificati dai geni HLA-A (ciascuno ereditato da un genitore), 2 dai geni HLA-B e 2 dai geni HLA-C. La combinazione di diversi tipi HLA contribuisce ad ampliare lo spettro di `antigeni` presentabili a livello individuale, mentre il polimorfismo dei geni HLA implica una notevole variabilità fra individui in termini di `antigeni` presentati: in altre parole, individui diversi tenderanno a presentare porzioni diverse del proteoma di un virus. Questo, congiuntamente al fatto che una grossa porzione dei `recettori` sono privati, rende ancora più radicale la diversità inter-individuale della risposta immunitaria a livello microscopico, pur nella straordinaria convergenza e stabilità dei suoi esiti.

Dati di sequenziamento

La caratterizzazione quantitativa dei processi di generazione e selezione (nonché della diversità dei repertori risultanti) è stata resa possibile dai progressi tecnologici nella produzione di dati di sequenziamento, ovvero *set* di dati che, a partire dalla catena di nucleotidi, forniscono la composizione in amminoacidi (la **sequenza**) sia dei `recettori` del repertorio immunitario sia delle migliaia di `antigeni` presentati dalle cellule in campioni biologici prelevati da sangue o tessuti.

Grazie anche alla raccolta sistematica di queste sequenze in banche dati specializzate, è possibile quindi iniziare a studiare insiemi di `antigeni` presentati dallo stesso HLA o repertori di `recettori` da diversi individui specifici verso lo stesso `antigene`, in numeri grandi a sufficienza da consentire un'analisi di tipo statistico. E proprio a livello statistico, emergono motivi ricorrenti fra `sequenze`, *pattern*, che riflettono dei vincoli, in termini di composizione in amminoacidi, imposti dalla conformazione strutturale e dalle affinità di legame (quantità, queste, per cui tuttavia la scala di disponibilità di dati sperimentali è ridotta rispetto alle sequenze).

Siccome produrre dati di sequenziamento in grandi quantità è divenuto sempre più efficiente e allo stesso tempo economico, si pone la necessità, nell'ambito dell'immunologia così come negli altri ambiti impattati da questo avanzamento tecnologico, di sviluppare approcci computazionali che permettano di sfruttare al massimo quest'informazione statistica per caratterizzare il comportamento del sistema immunitario e formulare predizioni a riguardo, ad esempio per scopi clinici.

Partire da osservazioni statistiche per svolger-

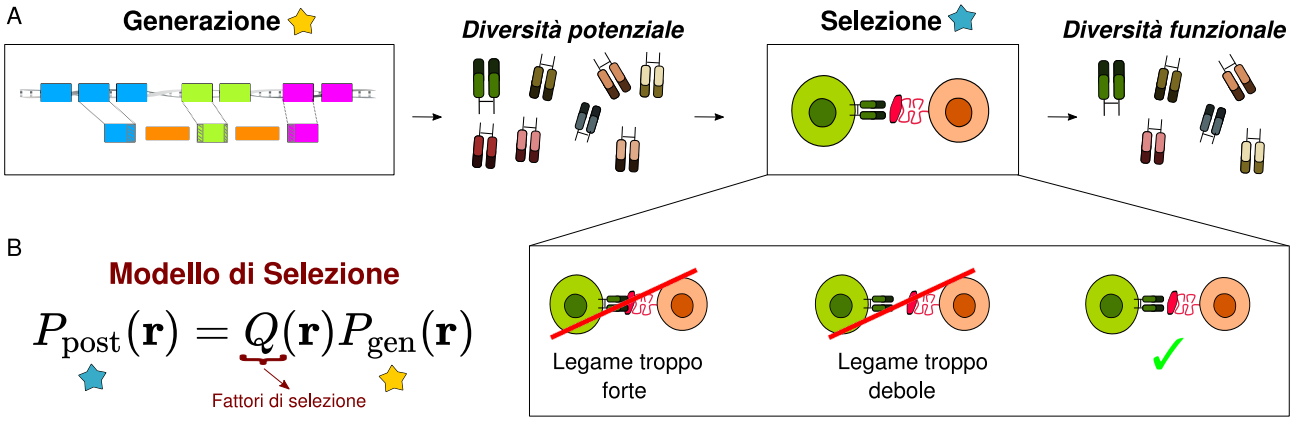

Figura 3: *A: Il processo di selezione timica agisce sui* `recettori` *generati dalla ricombinazione V(D)J, scartando quelli che non hanno proprietà di legame con gli* `antigeni` *idonee ad essere funzionali e riducendo così la diversità del repertorio risultante dal processo di generazione. B: Formula per la distribuzione di probabilità per il repertorio dei* `recettori` *post-selezione che specifica il modello di selezione proposto in [10, 11].*

ne un'analisi predittiva, la quale sostanzialmente richiede una componente di modellizzazione dei dati, costituisce un esempio di ciò che, in meccanica statistica, si chiama risolvere un **problema inverso**.

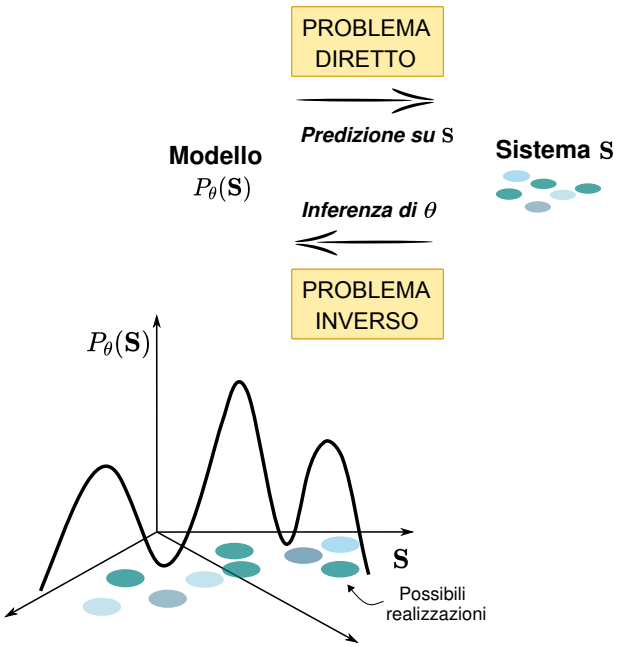

Figura 4: *Il problema inverso in meccanica statistica. In relazione alla modellizzazione della risposta immunitaria, le diverse realizzazioni del sistema sono i set di* `antigeni` *e di* `recettori` *dei* `linfociti` *T messi a disposizione dalle tecniche di sequenziamento high-throughput.*

La modellizzazione

Il problema inverso in meccanica statistica

In fisica, il problema diretto consiste nella predizione dello stato o della dinamica del sistema a partire da un modello che ne descrive il comportamento fisico in dipendenza di certi parametri noti (e fissati per esempio attraverso delle misure: dei tempi tipici di vita di una specie molecolare, delle energie libere di reazione, delle costanti di diffusione etc.).

Il problema inverso, invece, consiste nell'inferenza del valore di questi parametri una volta che sono date delle osservazioni empiriche (Fig. 4).

La logica di funzionamento del problema diretto e di quello inverso è di per sé generale in fisica; quando si parla di problema inverso in meccanica statistica, il modello è specificato da una distribuzione di probabilità (che assegna un valore di probabilità a ogni realizzazione del sistema), le predizioni sono formulate in termini di osservabili macroscopiche, statistiche, come possono essere la media e le correlazioni, e l'inferenza stessa parte da osservabili statistiche stimate su un insieme di realizzazioni del sistema che ne rappresentano un campionamento statistico.

Nel quadro della modellizzazione della risposta immunitaria, sono i repertori dei `recettori` o il complesso degli `antigeni` a poter essere pensati come degli insiemi delle molteplici realizzazioni possibili esplorate dal mondo vivente di uno stesso tipo di molecola biologica. È dunque

come se questo tipo di dati di sequenziamento desse accesso agli spazi di possibilità di quelle molecole, spazi dove però la naturale variabilità è vincolata, o meglio incanalata (per usare un termine chiave nella teorizzazione del vivente [5]), da pressioni evolutive, selettive e funzionali.

In questo contesto, il problema inverso è equivalente a dedurre la distribuzione di probabilità in questo spazio, una distribuzione che rispecchia i vincoli funzionali che hanno contribuito a plasmarlo (come quei vincoli di composizione che favoriscono alte affinità di legame biochimico).

A seconda del *set* di dati considerato come campionamento statistico dello spazio di interesse, si potranno ricostruire le distribuzioni di vari processi che regolano l'attività del sistema immunitario, per esempio la distribuzione di probabilità di presentazione di un `antigene` o del suo riconoscimento, da vedere essenzialmente come strumenti che quantificano la propensione di una sequenza verso le interazioni molecolari responsabili di questi processi.

La ricostruzione di una distribuzione di probabilità è un problema all'intersezione con la statistica, e che in particolare interpella nozioni e tecniche di inferenza statistica (o apprendimento statistico, ereditando un linguaggio più proprio dell'informatica). Il problema si riconduce infatti ad un compito tipico di inferenza statistica, in cui si assume la forma parametrica di un modello (qui, un modello specificato da una distribuzione di probabilità) ed i valori dei parametri vengono inferiti (appresi) nella maniera che meglio aderisce al dato empirico.

L'apprendimento consiste, a livello matematico, nell'aggiustare i parametri del modello, solitamente attraverso algoritmi iterativi, affinché i valori dei parametri appresi soddisfino qualche determinato criterio; nel caso dell'apprendimento di una distribuzione di probabilità, un criterio standard è la massimizzazione della probabilità specificata da quei parametri stimata sui dati. Questo criterio è noto come *maximum likelihood* ed, in altre parole, equivale alla prescrizione che i parametri appresi devono essere quelli che rendono massima l'evidenza del dato empirico sotto il modello statistico ipotizzato. Si tratta di un criterio di per sé attinto dall'ambito della statistica, ma di cui le tecniche sempre più avanzate di apprendimento statistico e automatico, anche di interesse per la fisica statistica, potenziano la messa in pratica.

Un vantaggio particolare dei modelli espressi da una distribuzione probabilità, specialmente nell'ottica della loro applicazione alle molecole biologiche, è la caratteristica di essere **generativi**, ovvero la capacità di generare, tramite campionamento Monte Carlo a partire dalla distribuzione inferita, dei dati sintetici che possiedono le stesse proprietà statistiche dei dati di partenza. La generazione così intesa, infatti, è strumentale sia all'esplorazione quantitativa di questi spazi (permettendo per esempio di stimarne l'entropia) sia al *design* di molecole artificiali con caratteristiche funzionali (una volta combinata con test sperimentali serrati).

Si noti che la stessa idea è al centro di approcci come *Direct Coupling Analysis*, basato sull'inferenza di una distribuzione di probabilità à la Potts (ovvero, la distribuzione di un modello di fisica statistica con interazioni a due corpi) per descrivere le famiglie di proteine[1]. Vari lavori negli ultimi dieci anni hanno dimostrato come questo approccio sia in grado di sfruttare l'informazione nelle correlazioni statistiche che emergono durante l'evoluzione per identificare i siti della sequenza che sono a contatto nella struttura tridimensionale e per predire interazioni specifiche fra proteine (si veda [6] per una rassegna).

Modelli di apprendimento statistico

È possibile allora mettere in moto l'ingranaggio del problema inverso sui dati di sequenziamento per studiare la risposta immunitaria?

Qualche paragrafo fa si è accennato a quattro processi fondamentali per la funzionalità del sistema immunitario adattativo: la generazione dei `recettori` attraverso la `ricombinazione V(D)J`, la `selezione timica`, la `presentazione dell'antigene` ed il suo riconoscimento. Per ciascuno di questi processi, descriverò ora alcuni modelli di recente pubblicazione incentrati sull'idea di risolvere un problema inverso di meccanica statistica attraverso tecniche di apprendimento statistico.

[1] Per `famiglia` qui si intende l'insieme delle varianti della stessa proteina in diverse specie legate dal punto di vista evolutivo.

Generazione dei recettori

Il modello di generazione dei `recettori`, dapprima proposto in [7] ed in seguito implementato in pacchetti `software` in [8, 9], si prefigge di ricostruire la distribuzione di probabilità degli eventi di ricombinazione V(D)J, in maniera tale che, dato una generico `recettore`, gli si possa assegnare un valore di probabilità di essere generato attraverso questi eventi.

A livello matematico, il modello si basa sul tradurre in un linguaggio probabilistico l'insieme di meccanismi stocastici in cui si articola la ricombinazione (ovvero la scelta dei segmenti VDJ, `inserzioni` e `delezioni` alle giunzioni) per definire la probabilità di un determinato scenario di ricombinazione (la $P(\texttt{scenario})$ riportata in Fig. 2B). Siccome vari scenari di ricombinazione possono dare luogo ad uno stesso `recettore` finale, la probabilità di un `recettore` di sequenza $\mathbf{r}$ di essere generato, $P_{\text{gen}}(\mathbf{r})$, può essere stimata come:

$$P_{\text{gen}}(\mathbf{r}) = \sum_{\text{scenari}} P(\text{scenario}),$$

dove la somma è intesa sugli scenari di ricombinazione passibili di aver generato la `sequenza` $\mathbf{r}$.

Il punto di partenza per l'apprendimento *data-driven* del modello sono dei *set* di sequenze che rappresentano il risultato puramente del processo di ricombinazione[2]. I parametri che specificano $P(\texttt{scenario})$ vengono appresi da *set* di dati di questo tipo attraverso un algoritmo iterativo di massimizzazione della *likelihood* del modello noto come *Expectation-Maximization*.

In questo contesto un modello probabilistico è uno strumento estremamente utile per quantificare la diversità risultante dal processo di generazione. Un primo indicatore di questa diversità è dato dall'entropia della distribuzione di probabilità inferita: per esempio, si è stimato [9] che lo spazio dei `recettori` dei `linfociti` T (catena β) abbia un'entropia di ~ 44 bits, la quale corrisponde a una dimensione effettiva (ossia come se la distribuzione fosse uniforme) di $\sim 10^{13}$ `sequenze` diverse.

Inoltre, l'inferenza di un modello relativamente semplice, che prevede dei termini di probabilità fattorizzati per i vari meccanismi coinvolti nella generazione (Fig. 2B), consente di dissezionarne i contributi alla diversità risultante. Per esempio, essendo le possibili combinazioni dei segmenti VDJ solo un migliaio, il maggiore contributo ad un processo di diversificazione che dà luogo a milioni di `recettori` deriva da `inserzioni` e `delezioni`; secondo le stime riportate in [7], questo contributo equivale circa all'83% dell'entropia degli eventi di ricombinazione.

Selezione dei recettori

La diversità risultante dal processo di generazione può essere considerata come la diversità potenziale dei `recettori` [3] ed il processo successivo di selezione timica ne comporta una netta riduzione, come quantificabile attraverso un modello probabilistico per tale selezione.

Il modello discusso negli articoli [10, 11] serve a descrivere la distribuzione statistica P_{post} delle sequenze post-selezione, ovvero delle sequenze che superano la selezione timica e che quindi si trovano, di fatto, nel repertorio immunitario individuale in condizioni normali.

Esempi di *set* di dati da cui può essere inferito provengono dal sequenziamento dei `recettori` contenuti in campioni di sangue. La scrittura del modello (vedi anche Fig. 3B) si basa sulla definizione di fattori di selezione $Q(\mathbf{r})$ tali che:

$$P_{\text{post}}(\mathbf{r}) = Q(\mathbf{r}) P_{\text{gen}}(\mathbf{r}), \qquad (1)$$

ovvero tali da incorporare gli effetti della selezione che agiscono, a livello statistico, sulla composizione di un `recettore` $\mathbf{r}$ in aggiunta ai vincoli di composizione imposti dal processo di generazione (e riassunti da P_{gen}).

Come per il modello di generazione, i parametri (qui i fattori Q) sono inferiti dai dati seguendo

[2] Dati di questo tipo sono forniti dalle cosiddette `sequenze` *out-of-frame*, ossia ricombinazioni del materiale genetico che risulterebbero in `recettori` non-funzionali. Tipicamente in questi casi la ricombinazione viene ripetuta a partire dal secondo cromosoma per produrre un `recettore` funzionale, ma il DNA del linfocita conserva comunque la sequenza ricombinata *out-of-frame*. In quanto non espresse come `recettori`, le sequenze *out-of-frame* non sono sottoposte ad alcuna selezione, restituendo un ritratto fedele del solo processo di generazione.

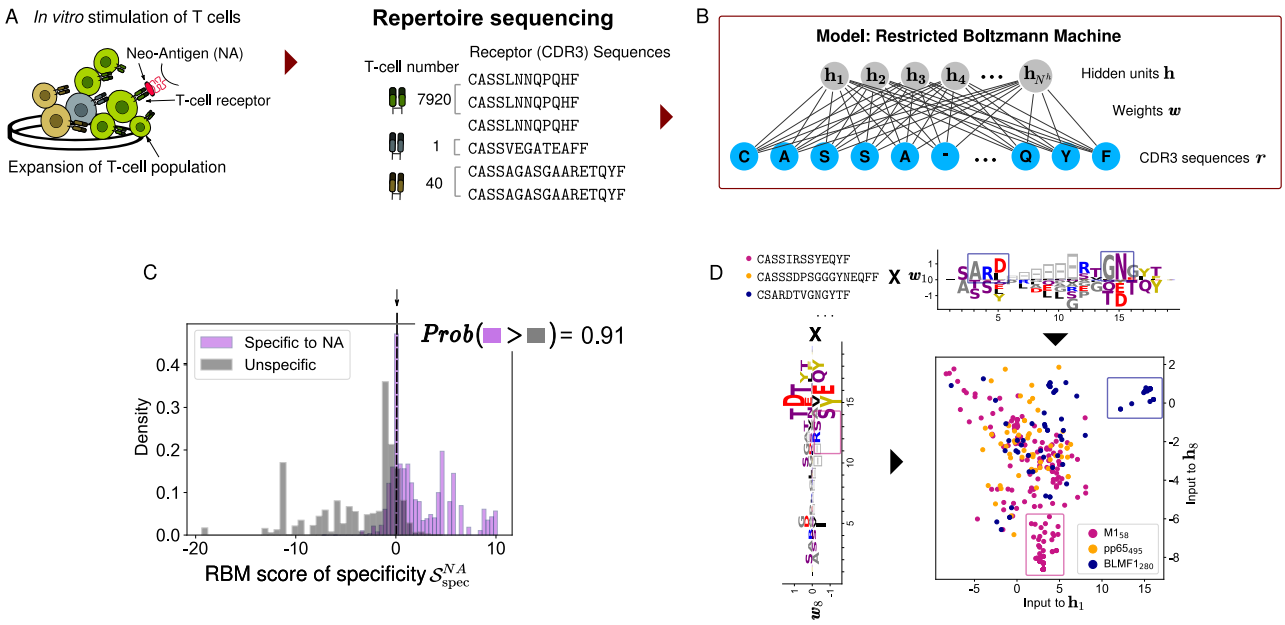

Figura 5: *Esperimenti di stimolazione in vitro dei* `linfociti` *T con* `neoantigeni` *specifici del paziente (effettuati da [16]) restituiscono set di dati in cui le abbondanze dei* `recettori` *sono rappresentative del grado di reattività dei* `linfociti` *corrispondenti (A) e da cui si può apprendere, attraverso la RBM (B), la probabilità di risposta ad un certo* `neoantigene` *NA, $P^{NA}(\mathbf{r})$. C: Abbiamo definito uno score di specificità di risposta al* `neoantigene` *NA come $\mathcal{S}^{NA}_{spec}(\mathbf{r}) = log\, P^{NA}(\mathbf{r}) - \max_a log\, P^a(\mathbf{r})$ (dove a nel secondo termine varia fra le stimolazioni* `antigeniche` *testate nello stesso paziente). I valori $\mathcal{S}^{NA}_{spec}$ consentono di discriminare i* `recettori` *che rispondono specificamente al* `neoantigene` *NA (valori positivi) da quelli che non vi sono specifici, come testimoniato dalla stima della probabilità (0.91) che il modello assegni uno score più alto ai* `recettori` *che effettivamente appartengono a popolazioni di* `linfociti` *T espanse. D: Modello RBM di riconoscimento appreso da un repertorio che raccoglie* `recettori` *dagli esperimenti di [18] specifici ad* `antigeni` *virali comuni ($M1_{58}$: virus dell'influenza, $pp65_{495}$: citomegalovirus, $BMLF1_{280}$: virus Epstein-Barr). Ogni input ad una unità nascosta raggruppa i dati in base a proprietà globali catturate dai pesi che comunicano con quella unità: qui ogni input riunisce in un cluster (indicato dal quadrato magenta o blu) i* `recettori` *che possiedono il gruppo funzionale di amminoacidi, anch'esso indicato da un quadrato magenta o blu, ravvisabile nei pesi corrispondenti visualizzati (ogni lettera sta per un amminoacido e la sua altezza per il valore del peso associato). I due diversi gruppi funzionali caratterizzano* `recettori` *che riconoscono due virus diversi (come segnalato dal colore). La figura è adattata da [12].*

il criterio di *maximum likelihood*, il quale è anche equivalente a richiedere che la probabilità del modello P_{post} riproduca nel modo migliore possibile la distribuzione statistica degli amminoacidi alle singole posizioni del ricettore osservata nei dati.

L'entropia associata alla distribuzione inferita, P_{post}, consente di stimare quantitativamente che la riduzione di diversità legata alla selezione porta dai $\sim$ 44 bits menzionati per il processo di generazione a $\sim$ 38 bits [10], corrispondente a una dimensione effettiva dello spazio dei `recettori` post-selezione di $\sim 10^{11}$ `sequenze`. Si tratta qui della diversità compatibile con la funzionalità dei `recettori`, una diversità esponenzialmente grande seppur notevolmente diminuita rispetto alla diversità potenziale messa in atto dalla generazione: la selezione delle `sequenze funzionali` corrisponde infatti all'eliminazione, principalmente, delle `sequenze rare`, cioè delle `sequenze` che hanno più bassa probabilità di essere generate ma che, per questa ragione, contribuiscono maggiormente alla diversità potenziale. Grosso modo, possiamo pensare al repertorio individuale di $\sim 10^8$ `recettori` come una porzione, estratta in maniera aleatoria, di questo spazio della diversità funzionale.

I fattori di selezione Q, inoltre, possono essere facilmente visualizzati, rilevando così le posizioni e/o i tipi di amminoacidi maggiormente affetti dal processo di selezione.

L'aspetto di visualizzazione è particolarmente vantaggioso quando si scelgono, come possibili *set* di dati di partenza, dei repertori di `recettori`

sottoposti a certi stimoli antigenici, dove i fattori di selezione contribuiscono ad identificare un arricchimento mirato di certi motivi di sequenza statisticamente associati alla risposta immunitaria a quell'`antigene`. Questo è uno dei risultati che mostriamo in [12], dove abbiamo utilizzato un modello con la struttura dell'eq. (1) alla stregua di quello che potremmo piuttosto chiamare un modello del riconoscimento specifico dell'`antigene`.

Riconoscimento specifico dell'antigene

È possibile mettere a punto dei modelli statistici dei repertori immunitari anche per predire la probabilità di riconoscimento specifico da parte dei `linfociti` T di `antigeni` di origine virale o tumorali, riconoscimento che avviene attraverso l'instaurazione di un legame biochimico specifico fra questi ultimi e i `recettori` di membrana dei `linfociti` T.

Nel lavoro [12] abbiamo sviluppato modelli di questo tipo non solo convertendo il modello di selezione a questo scopo (vedi sopra), bensì soprattutto ricorrendo ad un altro approccio, ideato nella comunità dell'apprendimento automatico [13, 14] e noto come *Restricted Boltzmann Machines* (che d'ora in poi indicherò con l'acronimo RBM). Tale approccio, fra l'altro, è stato già egregiamente presentato su questa rivista sia da Daniele Tantari che da Aurélien Decelle [15].

L'architettura di una RBM è riportata in Fig. 5B: per riassumerla brevemente, i singoli siti delle `sequenze` osservate che si intendono modellizzare (qui le `sequenze` del `recettore` $\mathbf{r}$) sono accoppiati ad un livello di unità nascoste $\mathbf{h}$ tramite un insieme di connessioni chiamate `pesi`.

A livello matematico, il modello è dunque specificato dalla distribuzione di probabilità congiunta $P(\mathbf{r},\mathbf{h})$ fra le sequenze osservate $\mathbf{r}$ e le unità nascoste $\mathbf{h}$ di cui la distribuzione di probabilità dei dati $P(\mathbf{r})$ è la distribuzione marginalizzata sulle unità nascoste:

$$P(\mathbf{r}) = \int \prod_\mu dh_\mu P(\mathbf{r},\mathbf{h}), \quad P(\mathbf{r},\mathbf{h}) \sim e^{-\mathcal{E}(\mathbf{r},\mathbf{h})},$$

$$\mathcal{E}(\mathbf{r},\mathbf{h}) = \underbrace{-\sum_i g_i(r_i)}_{\text{Campi locali}} + \underbrace{\sum_\mu \mathcal{U}_\mu(h_\mu)}_{\text{Potenziali nascosti}} \underbrace{-\sum_{i,\mu} h_\mu w_{i\mu}(r_i)}_{\text{Termine con i pesi}} . \tag{2}$$

In quest'ultima espressione l'indice i denota le unità osservate mentre l'indice μ le unità nascoste.

Il vantaggio legato all'introduzione di un livello di unità nascoste (che a buon diritto può essere visto un po' come un artificio, una convenzione del modello) è quello di rendere il modello particolarmente espressivo, pur mantenendolo trattabile dal punto di vista analitico: riscrivere una distribuzione $P(\mathbf{r})$ in termini di una distribuzione congiunta consente infatti di includere delle correlazioni fra i siti delle sequenze $\mathbf{r}$ anche di ordine superiore al secondo, senza una proliferazione di termini di interazione a molti corpi e dei parametri associati.

In maniera analoga ai modelli di generazione e selezione discussi finora, tutti i parametri del modello scritti nell'eq. (2) (i `pesi`, i campi locali sulle unità osservate, i parametri che specificano i potenziali sulle unità nascoste) vengono appresi a partire dai dati ottemperando al criterio di *maximum likelihood*.

I dati utili a costruire il modello RBM di riconoscimento provengono dal sequenziamento dei `recettori` in campioni di sangue dove una risposta dei `linfociti` viene stimolata *in vitro* attraverso l'esposizione di un `antigene` selezionato. Repertori di `recettori` ottenuti in questo modo possono essere visti come un campionamento del processo di espansione delle popolazioni di `linfociti` T che segue al riconoscimento dell'`antigene` (Fig. 1): i diversi `recettori` vi compaiono dunque con abbondanze diverse, e le abbondanze maggiori sostanzialmente segnalano le popolazioni più espanse che dominano la risposta immunitaria (Fig. 5A).

Da *set* di dati con questa struttura, abbiamo appreso attraverso la RBM una distribuzione di probabilità $P(\mathbf{r})$ interpretabile come la probabilità che il linfocita col `recettore` $\mathbf{r}$ riconosca l'`antigene` considerato nell'esperimento.

Muniti di queste probabilità per diversi `antigeni` testati in diversi individui, le possiamo confrontare al fine di identificare quali `recettori` sono reattivi specificamente ad un

Che cos'è l'immunoterapia?

Per **immunoterapia** si intendono tutti quei trattamenti oncologici mirati a stimolare nel paziente una risposta immunitaria contro il tumore. Possono essere suddivisi in tre macro-tipologie:

- *Immune Checkpoint Inhibitors*: farmaci che inibiscono le molecole impegnate a frenare i `linfociti` T, sbloccandone così l'azione anti-tumorale. La loro scoperta è valsa il Nobel per la Medicina 2018, assegnato congiuntamente a James P. Allison e Tasuku Honjo.
- Terapie basate sulle cellule CAR-T (dall'inglese *Chimeric Antigen Receptor T cell*): approccio in cui `linfociti` T sono direttamente prelevati dal paziente, modificati in laboratorio per accrescerne la specificità di riconoscimento delle cellule tumorali e ri-trasferiti al paziente.
- Vaccini anti-cancro: terapie basate sulla somministrazione di `neoantigeni` (`antigeni` contenenti mutazioni acquisite durante la progressione tumorale) in maniera da allenare una risposta immunitaria contro le cellule tumorali che li presentano.

In parte queste terapie sono approvate per l'impiego nella prassi clinica per alcuni tipi di tumore (come il melanoma), in parte sono in fase di sperimentazione. Per una descrizione più dettagliata si rimanda a [17].

determinato `antigene` e non ad altri (Fig. 5C). In particolare, ci siamo concentrati sulla modellizzazione quantitativa del riconoscimento dei cosiddetti `neoantigeni`, ovvero `antigeni` provenienti da proteine proprie dell'organismo umano ma riconoscibili come estranei in quanto contengono mutazioni occorse durante la progressione tumorale; i dati sperimentali a nostra disposizione (dal lavoro [16]) descrivono la risposta in pazienti affetti da tumore al pancreas.

Lo studio delle reazioni immunitarie ai `neoantigeni` si inserisce in un contesto di ricerca più ampio, che mira a costruire un quadro di comprensione della risposta immunitaria contro le cellule tumorali messa in campo dai `linfociti` T ed in parallelo a supportarne il trasferimento nell'ambito clinico dell'immunoterapia (vedi riquadro).

A livello di predizioni del modello, è possibile osservare che i parametri inferiti dell'RBM (in particolare i pesi) rilevano dei motivi di sequenza potenzialmente funzionali, ed in generale è ragionevole ipotizzare che tali motivi identifichino certi gruppi di amminoacidi che contribuiscono a stabilire il legame biochimico con l'`antigene`, fornendo così degli input precisi per dei test sperimentali. Un'evidenza preliminare: i motivi di sequenza indicati in Fig. 5D, ottenuti apprendendo un modello RBM di riconoscimento su *set* di `recettori` specifici ad `antigeni` virali estremamente comuni come l'influenza, sono validabili in questa interpretazione funzionale a partire dalle strutture del complesso `antigene-recettore`, come mostrato in [18].

In aggiunta, la proiezione delle sequenze $\mathbf{r}$ sui pesi che le collegano con una certa unità nascosta (il cosiddetto input di una unità nascosta, Fig. 5D) disegna le coordinate di un nuovo spazio di rappresentazione che "comprime" i dati amplificando i *pattern* e le regolarità rilevanti all'apprendimento delle loro proprietà statistiche; nel caso delle proteine, tali rappresentazioni interne sono riconducibili a caratteristiche funzionali, strutturali, filogenetiche condivise da sottogruppi delle sequenze modellizzate[3] [19]. L'idea alla base, ben illustrata proprio dall'applicazione ai `recettori` dei `linfociti` T (Fig. 5D), è che gli input alle unità nascoste separano gruppi di `recettori` accomunati dai motivi di sequenza

[3] Il numero delle unità nascoste è uno dei cosiddetti "iperparametri" del modello, in quanto pertinenti alla sua struttura globale, e viene fissato con procedure standard di validazione incrociata in cui si effettua l'apprendimento su una porzione dei dati per poi valutarne la performance sulla porzione rimanente. Intuitivamente, allora, la ricerca iperparametrica del numero di unità nascoste può essere pensata come una ricerca empirica del numero più adeguato di *pattern*, sottotipologie o caratteristiche per rappresentare i dati in considerazione.

rilevati dai pesi ed, in virtù dell'interpretazione funzionale di questi motivi in termini di specificità di legame con l'antigene, permettono dunque di ben separare gruppi di recettori con specificità di riconoscimento verso virus diversi.

L'identificazione di motivi di sequenza potenzialmente funzionali mostra come un modello probabilistico del riconoscimento possa non solo fungere da strumento per individuare quali recettori rispondono ad un dato antigene, ma possa anche contribuire ad isolare, almeno parzialmente, quali proprietà a livello biochimico lo rendono in grado di rispondervi in modo specifico. Modelli statistici che affrontano in maniera più puntuale questa domanda, tuttavia, dovrebbero descrivere l'interazione specifica fra i siti dell'antigene e quelli del recettore: la possibilità di costruire modelli statistici di questo tipo è resa sempre più concreta dalla crescita costante di dati sperimentali, che sopperirà gradualmente anche alla mancanza di un campionamento *high-throughput* e combinato di antigeni e dei recettori che li riconoscono, ovvero un campionamento del processo di interazione antigene-recettore.

Presentazione dell'antigene

Le RBM sono un approccio che consente di modellizzare a livello probabilistico anche il processo complementare al riconoscimento dell'antigene, ossia la presentazione di quest'ultimo da parte di una determinata molecola HLA.

Io ed i miei collaboratori abbiamo utilizzato le RBM con questo scopo in [20], partendo da *set* di dati sperimentali che mappano grandi quantità di antigeni presentati dalle proteine HLA in un certo campione biologico (un tessuto per esempio).

Da questi dati abbiamo appreso una distribuzione di probabilità, parametrizzata come nell'eq. (2), che può essere interpretata come la distribuzione di probabilità del processo di presentazione, ovvero una distribuzione che incorpora i vincoli, a livello di composizione in amminoacidi, che rendono un antigene presentabile. Questo modello RBM è poi divenuto il nucleo di funzionamento di un metodo che abbiamo denominato RBM-MHC (dove MHC è un altro acronimo per indicare i complessi proteici HLA [1]).

Si tratta di un metodo sia per predire, assegnando uno *score* probabilistico di presentazione attraverso la distribuzione di probabilità inferita, quali porzioni di proteine, virali o tumorali, sono presentate al sistema immunitario (Fig. 6A-D), sia per predire il tipo di proteina HLA che più plausibilmente presenta un certo antigene[4] (Fig. 6C-E).

A questo proposito, nello spazio di rappresentazione della RBM (Fig. 6B), gli antigeni si organizzano in *cluster* (gruppi) sulla base della composizione biochimica, e, siccome le proprietà di legame con la proteina HLA dipendono dalla presenza di certi gruppi funzionali di amminoacidi, *cluster* differenti corrispondono ad antigeni che si legano in maniera specifica a differenti proteine HLA. Lo scopo è stato allora sfruttare questa rappresentazione dei dati interna al modello per predire quale proteina HLA presenta un certo antigene, costruendo, sempre tramite l'apprendimento automatico, un classificatore della tipologia HLA (Fig. 6C).

La struttura in *cluster* della rappresentazione compressa dei dati elaborata dalla RBM, che già separa gli antigeni in base alla specificità di legame con l'HLA, fa sì che questo classificatore possa essere appreso utilizzando solo una piccola percentuale di antigeni, di tipologia HLA nota dalle banche dati, per poi essere in grado di associare accuratamente gli antigeni restanti alla tipologia HLA corrispondente (Fig. 6E). Tale approccio è dunque utile soprattutto per la caratterizzazione di tutti quei *set* di dati di recente produzione (per esempio in test clinici), per cui è possibile recuperare l'annotazione relativa alla specificità HLA dalle banche dati solo in una minoranza di casi.

Predizioni computazionali sulla presentazione vengono tipicamente sfruttate nello sviluppo dei vaccini, più specificatamente per la selezione iniziale delle regioni del proteoma di un virus presentabili dagli HLA più comuni nella popolazione umana, in modo da individuare un *set* li-

[4] Questa infatti è un'informazione che tipicamente non è nota quando un antigene viene rilevato sperimentalmente, ciò che è noto è il *set* dei 6 tipi di HLA posseduti a livello genetico dall'individuo da cui il campione è stato prelevato.

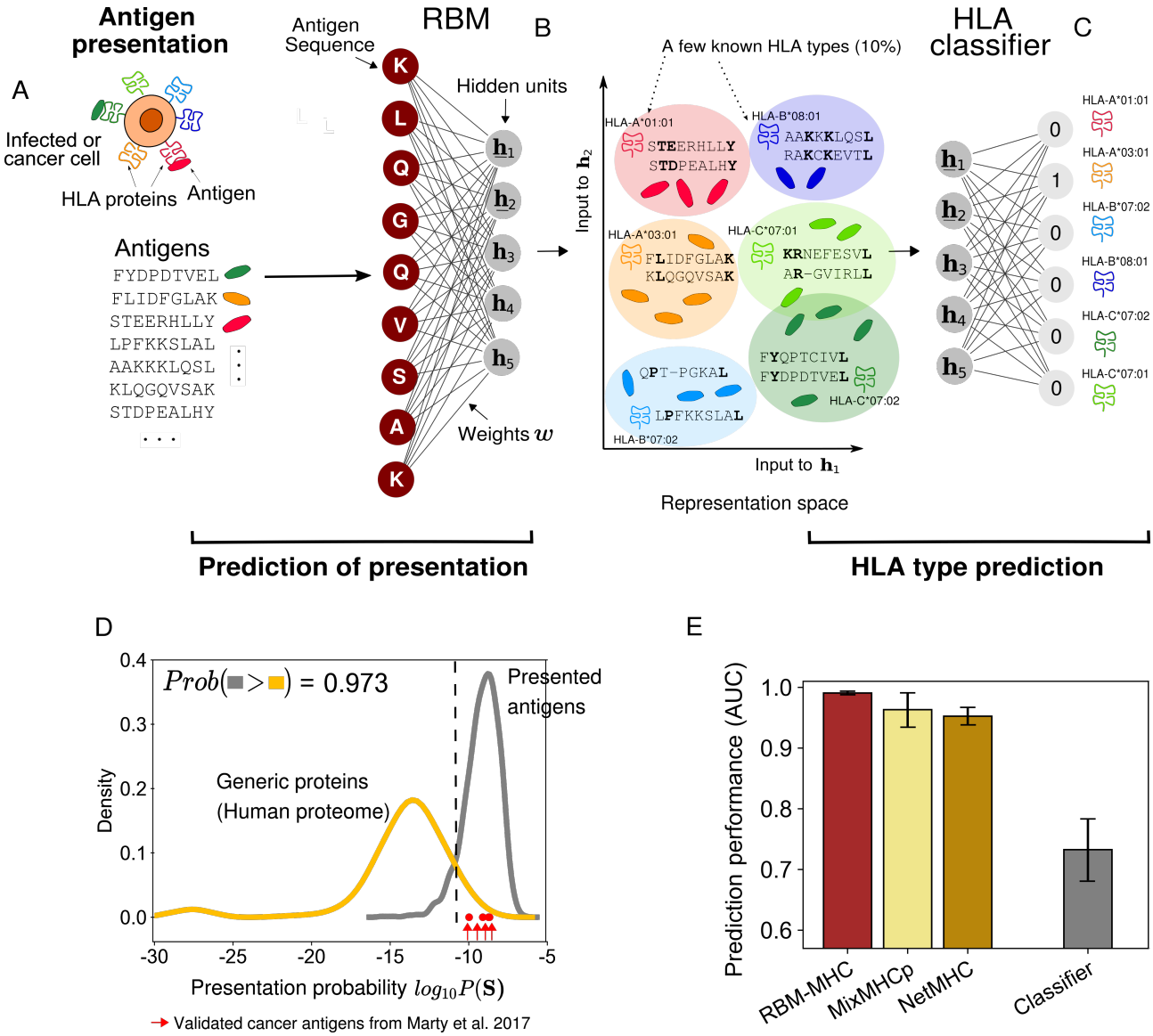

Figura 6: *Struttura del metodo RBM-MHC per la predizione della presentazione degli* `antigeni` *da parte di specifici HLA (A). La predizione di quali* `antigeni` *sono presentabili è effettuata assegnando la probabilità di presentazione inferita tramite una RBM (B): il pannello D mostra che questa quantità è capace di discriminare* `antigeni` *sperimentalmente rilevati come presentati (come degli* `antigeni` *tumorali, marcati in rosso) e generici frammenti di proteine che generalmente non sono presentabili. La qualità della predizione è misurata dalla probabilità che il modello assegni uno score probabilistico più alto agli* `antigeni` *effettivamente presentati (0.97). In un secondo momento, il metodo apprende un classificatore (C), in modo da sfruttare la struttura a cluster dello spazio di rappresentazione della RBM per predire il tipo di HLA che presenta ogni* `antigene`*. Anche quando le associazioni* `antigene`*-HLA a disposizione durante l'apprendimento coprono solo una piccola percentuale dell'intero set di* `antigeni`*, la predizione finale è estremamente accurata (E), del tutto paragonabile allo stato dell'arte (fornito da algoritmi come NetMHC [27]) e decisamente più alta di un classificatore appreso dalla stessa piccola porzione di associazioni note a partire direttamente dalle sequenze (e non dalla loro rappresentazione nello spazio degli input alle unità nascoste). La figura è adattata da [20].*

mitato di potenziali *target* antigenici da validare sperimentalmente in fasi successive.

Un ruolo simile è quello svolto nell'ambito dell'`immunoterapia` (vedi riquadro) ed in particolare nella cosiddetta *neoantigen discovery*, ovvero per l'identificazione dei `neoantigeni`, specifici al paziente, suscettibili di scatenare una risposta immunitaria da parte dello stesso, un passaggio chiave nello sviluppo di vaccini anticancro personalizzati. In questo contesto, metodi computazionali le cui predizioni possono essere formulate solo a partire dalla sequenza sono di estrema importanza: essi consentono infatti di filtrare speditamente quello che tipicamente è un enorme numero di candidati `neoantigeni` (provenienti da tutti i possibili frammenti di proteine contenenti delle mutazioni), andando a selezionare solo i frammenti che hanno un'alta probabilità di essere presentati dalle proteine HLA del paziente [21].

Per sviluppare procedure di *neoantigen discovery* realmente efficaci, la predizione della presentazione deve essere complementata da una stima del potenziale di riconoscimento dei `neoantigeni` da parte dei `linfociti` T del repertorio individuale. Identificare le proprietà degli `antigeni` che ne determinano il potenziale di riconoscimento e la predizione quantitativa di quest'ultimo sono tuttavia problemi assolutamente aperti [22], la cui soluzione aggiungerà un mattone fondamentale sia alla nostra comprensione teorica della risposta immunitaria sia ai nascenti approcci di medicina personalizzata.

Nel panorama del *Machine Learning*

Gli approcci *data-driven* fin qui descritti sono caratterizzati da strutture relativamente semplici, mappabili in modelli di meccanica statistica; l'impiego che ne ho delineato è essenzialmente quello di individuare dei *pattern* statistici in grandi moli di dati e ricavarne predizioni. Questo è esattamente uno di quei problemi computazionali che costituiscono il campo di applicazione classico di svariate tecniche di apprendimento automatico (o *machine learning*), come le reti neurali e, più di recente, le reti organizzate su svariati livelli dette profonde (*deep learning*[5]).

Con la rivoluzione apportata nell'ultima decade soprattutto dal *deep learning*, approcci di apprendimento automatico sono utilizzati in maniera crescente anche per affrontare i problemi di immunologia computazionale di cui si è discusso. Ne menziono solo alcuni a scopo illustrativo: modelli basati sull'architettura profonda del *Variational Autoencoder* [23], proposti come un'alternativa per ricostruire la distribuzione di probabilità del processo di generazione (senza però incorporarne esplicitamente i meccanismi noti); soNNia [24], una versione *deep* del modello di selezione qui presentato; classificatori dei `recettori` per specificità di risposta verso un determinato insieme di `antigeni` (come TCRex [25] e DeepTCR [26]); NetMHC [27], il metodo di più ampia diffusione per predire la presentazione dell'`antigene` da parte di una certa molecola HLA, il quale è strutturato come una rete neurale appresa in modo supervisionato a partire dall'informazione sull'affinità di legame degli `antigeni` all'HLA corrispondente.

Per contrasto con le architetture *deep*, la tipologia di apprendimento eseguito dalle RBM viene chiamata *shallow learning* (letteralmente, apprendimento superficiale), come è evidente dalla sua rappresentazione grafica, che prevede un solo livello nascosto (Fig. 5B). Chiaramente le architetture *deep*, combinando più livelli che elaborano l'informazione attraverso trasformazioni non lineari successive, sono in grado di specificare funzioni arbitrariamente complesse e di realizzare dei fit statistici dei dati sofisticati, capaci per esempio di individuare *pattern* complessi e di riprodurre proprietà statistiche che non si limitano alle sole medie o correlazioni a due punti.

C'è qualche vantaggio che dovrebbe motivarci a priviliegiare anche degli approcci *shallow*? Modelli dalla struttura più semplice sono tipicamente modelli più parsimoniosi in termini di parametri, proprio perché specificati da un numero inferiore di questi ultimi. Da un lato, modelli parsimoniosi consentono di mantenere sotto controllo meglio il problema dell'*overfitting*, che porta il modello appreso a conformarsi troppo dettagliatamente sul particolare *set* di dati a disposizione, perdendo potere di generalizzazione su dati nuovi. Questo problema è indotto dalla

[5] Si vedano, per delle spiegazioni più esaustive, gli articoli per esempio di Guido Sanguinetti, Carlo Lucibello e Giorgio Buttazzo sul recente numero di Ithaca dedicato all'intelligenza artificiale [15].

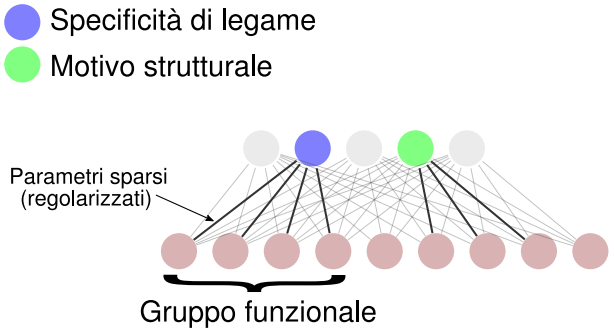

Figura 7: *Rappresentazione stilizzata di un esempio dell'effetto di una regolarizzazione che favorisce la sparsità: valori significativamente diversi da zero dei pesi (colore scuro) si concentrano su alcuni siti della sequenza, per esempio quelli in cui compare un gruppo funzionale di amminoacidi che ne determina la specificità di legame con altre molecole. L'unità nascosta connessa a quei pesi può essere immediatamente interpretata come l'unità che cattura questa proprietà e, ricevendo input di valore diverso dalle sequenze con varianti distinte del gruppo funzionale, aiuta a discriminare diverse specificità di legame. Similmente altre unità nascoste rendono manifeste altre proprietà, come la presenza di diversi motivi strutturali.*

condizione di inferire un numero troppo elevato di parametri a partire dall'informazione di un *set* di dati molto limitato (come è spesso il caso in biologia, soprattutto quando i dati sono prodotti in esperimenti gravosi sia dal punto di vista dei costi che del tempo impiegato); modelli con un numero modesto di parametri sono allora meno affetti da questo problema, a cui invece è necessario fare particolare attenzione con le architetture *deep*, tipicamente sovraparametrizzate.

Dall'altro lato, modelli parsimoniosi equivalgono sia a modelli i cui parametri possono essere ispezionati visualmente sia spesso a modelli più trattabili a livello matematico, ovvero studiabili tramite calcoli analitici: questi sono aspetti che favoriscono la cosiddetta interpretabilità del modello, un'altra istanza che può e deve legittimamente intervenire nella selezione del modello.

L'interpretabilità, nell'ambito dell'apprendimento automatico, non è un concetto con un'accezione univoca, potendosi riferire sia all'intellegibilità e alla trasparenza dell'algoritmo a livello di funzionamento sia al grado di spiegabilità del processo modellizzato e delle predizioni formulate sulla base del modello appreso [28]. La scelta di modelli trattabili e parsimoniosi è funzionale a questa seconda accezione (la spiegabilità dell'esito dell'apprendimento come modello del fenomeno), la quale può essere ulteriormente supportata da opportune scelte di termini di **regolarizzazione**.

I termini di regolarizzazione sono dei termini che vengono introdotti nell'inferenza per controllare i valori e le interdipendenze dei parametri durante il loro apprendimento; si tratta di termini particolarmente necessari ad evitare sia l'*overfitting* che possibili divergenze nei valori inferiti quando si ha un numero limitato di dati, permettendo allo stesso tempo di regolare il tipo ed il regime di rappresentazione fornita dai parametri. Per esempio, si può optare per una regolarizzazione che favorisca la sparsità dei parametri, ovvero una distribuzione disomogenea per cui i parametri significativamente diversi da zero si concentrano solo fra alcune unità dell'RBM e sono così più facilmente visualizzabili (Fig. 7). Una tale regolarizzazione ha anche il beneficio di rendere più separabile il contributo di alcuni sottogruppi di unità alla rappresentazione interna dei dati resa possibile dal livello nascosto, la quale potrà così essere interpretata in termini, per esempio, di proprietà biofisiche (si veda Fig. 7 e [19] per degli esempi con le proteine). È stato sistematicamente studiato [29] come le regolarizzazioni giochino un ruolo fondamentale per esempio nella transizione fra il regime di rappresentazione per prototipi e quella composizionale (ovvero per caratteristiche separate dei dati), descritta anche da Daniele Tantari su questa rivista [15]. Tali rappresentazioni per tipi o per proprietà condivise, a cui si è già accennato in precedenza, comportano una riduzione della dimensionalità dello spazio di modellizzazione dei dati, la quale può agevolare la scoperta di nuove regolarità ed associazioni, a loro volta da sfruttare per formulare nuove predizioni (un po' come la predizione della tipologia HLA illustrata in Fig. 6B-C); non è un caso, infatti, che ormai esistano numerosi algoritmi concepiti prima di tutto per la riduzione di dimensionalità dei dati (t-SNE, UMAP, DBSCAN etc.).

Domande aperte e limitazioni, ovvero spazi per la ricerca futura

L'enorme disponibilità di dati di un certo tipo (le `sequenze`) ha aperto nuove possibilità di modellizzazione a determinati livelli di descrizione, soprattutto molecolari. Le predizioni dei modelli ottenuti sono dunque basate sugli osservabili a cui questi dati garantiscono accesso: da una parte, le regolarità a livello di composizione in amminoacidi, indicative per esempio della propensione a stabilire legami biochimici `recettore-antigene`, e dall'altra le abbondanze delle singole sequenze, che consentono per esempio di quantificare il grado di espansione di alcune popolazioni di `linfociti` T.

Tuttavia una serie ulteriore di eventi, a cui pertengono modalità di misura e di descrizione matematica differenti, concorre ad indurre l'attivazione dei `linfociti` [3]: il tempo tipico di durata del legame chimico, il *clustering* dei `recettori`, la riorganizzazione dinamica della membrana dei `linfociti` con la formazione della sinapsi immunologica, modificazioni conformazionali, reazioni chimiche di *signaling* innescate dal legame `antigene-recettore` ecc. Il passaggio dal riconoscimento molecolare alla risposta è poi un effetto collettivo, di popolazione, sensibile al contesto dato dalle altre cellule del sistema immunitario.

Pertanto, la predizione basata sulle sequenze, seppur importantissima per estrarre informazioni dai dati disponibili in grande quantità, è intrinsecamente incompleta e deve allora, a maggior ragione, essere specificata a livello probabilistico.

Approcci che siano in grado di combinare diversi livelli di descrizione, integrando dati di diversa natura (strutture, affinità di legame oltre alle sequenze) costituiscono un obiettivo di comune interesse nella comunità allargata fra immunologia computazionale e biofisica statistica.

Certamente una domanda aperta, nonché di formidabile difficoltà dal punto di vista della scrittura dei modelli, è come scale globali e locali si influenzino a vicenda nel determinare la risposta immunitaria.

Anche dal punto di vista dell'impiego dell'apprendimento statistico nella modellizzazione del biologico vi sono da sottolineare delle limitazioni, sia teoriche e che pratiche, su cui parte del lavoro futuro potrebbe focalizzarsi. In questa modellizzazione statistica, come in ogni sforzo di modellizzazione, è implicita una maniera di rappresentare le variabili (qui le molecole biologiche) basata su delle assunzioni che apportano un vantaggio operativo ma non sempre completamente giustificate dal punto di vista dei fondamenti teorici.

Per esempio, pensiamo alle molecole biologiche come realizzazioni equivalenti ed indipendenti, quando invece, ad esempio per le famiglie di proteine, l'assunzione di indipendenza è particolarmente delicata siccome le proteine sono legate da relazioni filogenetiche (e correggere questi bias richiede soluzioni *ad hoc*); o ancora, le pensiamo in uno spazio statico, come se l'evoluzione avesse raggiunto l'equilibrio invece che essere un processo dinamico, continuamente in atto e plausibilmente fuori dall'equilibrio.

Similmente, modellizzare un intero spazio di possibili realizzazioni con i *set* di dati disponibili è un'approssimazione, soprattutto se guardiamo a spazi popolati da realizzazioni di un'enorme diversità, e con una distribuzione statistica eterogenea, come i `recettori` immunitari o gli `antigeni`.

Per quanto il quantitativo di dati di sequenziamento a disposizione cresca continuamente, dobbiamo per ora limitarci a pensarlo come un campionamento sì estremamente informativo ma sparso, non-esaustivo dello spazio intero.

Tecniche di evoluzione diretta [30] e di *deep mutational scanning* [31], che misurano sistematicamente gli effetti sulle proprietà biochimiche di legame di singole mutazioni, consentono dei campionamenti più densi dello spazio funzionale delle molecole biologiche, fornendo dunque la materia prima per costruire modelli con una più alta risoluzione sui dettagli molecolari del riconoscimento da parte dei `linfociti` e della loro conseguente attivazione.

Conclusione

In questo articolo ho descritto a grandi linee come si innesca, a livello molecolare, la risposta del sistema immunitario contro agenti patogeni e cellule tumorali, sottolineando un'interessante convivenza fra diversità e specificità che

pone sfide complesse sia di concettualizzazione che di formalizzazione matematica di modelli esplicativi per tali dinamiche.

In questo contesto, mi sono concentrata su schemi di modellizzazione che si innestano sulla maniera di pensare della meccanica statistica e che allo stesso tempo fanno ricorso all'apparato di tecniche dell'apprendimento statistico.

L'idea cardine consiste nel pensare i repertori immunitari come ipotetici spazi di possibilità fortemente vincolati da pressioni funzionali, e nel cercare dei modelli che realizzino una rappresentazione intellegibile di queste pressioni, estrapolando anche delle predizioni, attraverso l'inferenza delle distribuzioni di probabilità su questi spazi.

Si tratta dunque di approcci con un paradigma operativo ibrido, in quanto finalizzati ad elaborare modelli teorici ma con una forte componente induttiva ed empirica, derivante dall'aver incorporato l'informazione dei dati disponibili attraverso l'apprendimento statistico[6].

Gli esempi che ho fornito sono dei modelli quantitativi per quattro processi alla base della capacità di risposta dei `linfociti`: la generazione dei `recettori`, la loro selezione, il riconoscimento degli `antigeni` da essi implementato e la presentazione degli `antigeni`.

Nel discutere questi esempi, ho tentato di rendere chiaro infine che si tratta di approcci interdisciplinari non solo nel metodo ma anche nell'esito, in quanto hanno le potenzialità per confrontarsi con problemi di rilievo sia per l'immunologia che per la bioinformatica e la medicina più in generale.

Ringraziamenti

Ringrazio i miei mentori su questi argomenti, Simona Cocco, Rémi Monasson, Thierry Mora, Aleksandra M. Walczak, Vinod P. Balachandran e Benjamin D. Greenbaum, e tutte le persone da cui ho avuto modo di imparare sia nell'équipe "Physique Statistique et Inférence pour la Biologie" ad École Normale Supérieure che nella collaborazione "Computational Deconstruction of Neoantigen-TCR Degeneracy for Cancer Immunotherapy" supportata da Stand Up to Cancer - Lustgarten Foundation. Ringrazio in particolare Silvia Grigolon, Cosimo Lupo e Sara Torrenzieri per i loro preziosi commenti sul manoscritto.

[6] Se vogliamo, questo è proprio il paradigma operativo dell'ape nella similitudine di Francis Bacon riportata nell'*incipit*, la quale raccoglie materiale dal mondo circostante per poi trasformarlo con i propri strumenti e secondo i propri fini.

[1] L. Sompayrac, *How the Immune System Works*, 4th ed. Wiley-Blackwell, Chichester, West Sussex; Hoboken, NJ (2012).

[2] F.M. Burnet, *A modification of Jerne's theory of antibody production using the concept of clonal selection*, Aust. J. Sci., 20 (1957).

[3] G. Altan-Bonnet, T. Mora, A.M. Walczak, *Quantitative Immunology for Physicists*, Phys. Rep., 849 (2020).

[4] D. Mason, *A Very High Level of Crossreactivity Is an Essential Feature of the T-Cell Receptor*, Immunol. Today, 19 (1998) 9.

[5] G. Longo, M. Montévil, *Extended criticality, phase spaces and enablement in biology*, Chaos, Solitons & Fractals, 55 (2013).

[6] S. Cocco et al., *Inverse Statistical Physics of Protein Sequences: A Key Issues Review*, Rep. Prog. Phys., 81 (2018) 3.

[7] A. Murugan et al., *Statistical inference of the generation probability of T-cell receptors from sequence repertoires*, Proc. Natl. Acad. Sci. USA, 109 (2012) 40.

[8] Q. Marcou, T. Mora, A.M. Walczak, *High-throughput immune repertoire analysis with IGoR*, Nat. Commun., 9 (2018) 1.

[9] Z. Sethna et al., *OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs*, Bioinformatics, 35 (2019).

[10] Y. Elhanati et al., *Quantifying selection in immune receptor repertoires*, Proc. Natl. Acad. Sci. USA, 111 (2014) 27.

[11] Z. Sethna et al., *Population Variability in the Generation and Selection of T-Cell Repertoires*, PLoS Comput. Biol., 16 (2020) 12.

[12] B. Bravi et al., *Probing T-cell response by sequence-based probabilistic modeling*, PLoS Comput. Biol., 17 (2021) 9.

[13] P. Smolensky, *Information Processing in Dynamical Systems: Foundations of Harmony Theory*. In: D.E. Rumelhart, J.L. McLelland, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, Massachusetts (1986).

[14] G.E. Hinton, *Training Products of Experts by Minimizing Contrastive Divergence*, Neural Comput., 14 (2002) 8.

[15] *Ithaca: Viaggio nella Scienza. Intelligenza Artificiale*, XVI (2020)

[16] V.P. Balachandran et al., *Identification of Unique Neoantigen Qualities in Long-Term Survivors of Pancreatic Cancer*, Nature, 551 (2017) 7681.

[17] A.D. Waldman, J.M. Fritz, M. Lenardo, *A guide to cancer immunotherapy: from T cell basic science to clinical practice*, Nat. Rev. Immunol., 20 (2020) 11.

[18] P. Dash et al., *Quantifiable Predictive Features Define Epitope-Specific T Cell Receptor Repertoires*, Nature, 547 (2017) 7661.

[19] J. Tubiana, S. Cocco, R. Monasson, *Learning Protein Constitutive Motifs from Sequence Data*, eLife, 8 (2019).

[20] B. Bravi et al., *RBM-MHC: A Semi-Supervised Machine-Learning Method for Sample-Specific Prediction of Antigen Presentation by HLA-I Alleles*, Cell Syst., 12 (2021) 2.

[21] V. Roudko, B.D. Greenbaum, N. Bhardwaj, *Computational Prediction and Validation of Tumor-Associated Neoantigens*, Front. Immunol., 11 (2020) 27.

[22] D.K. Wells et al., *Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction*, Cell, 183 (2020) 3.

[23] K. Davidsen et al., *Deep Generative Models for T Cell Receptor Protein Sequences*, eLife, 8 (2019).

[24] G. Isacchini et al., *Deep generative selection models of T and B cell receptor repertoires with soNNia*, Proc. Natl. Acad. Sci. USA, 118 (2021) 14.

[25] S. Gielis et al., *Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires*, Front. Immunol., 10 (2019).

[26] J. Sidhom et al., *DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires*, Nat. Commun., 12 (2021) 1.

[27] M. Andreatta, M. Nielsen, *Gapped sequence alignment using artificial neural networks: Application to the MHC class I system*, Bioinformatics, 32 (2015).

[28] Z. Lipton, *The Mythos of Model Interpretability*, 2016 ICML Workshop on Human Interpretability in Machine Learning, New York (2016).

[29] J. Tubiana, S. Cocco, R. Monasson, *Learning compositional representations of interacting systems with Restricted Boltzmann Machines: Comparative study of lattice proteins*, Neural Comput., 31 (2019) 8.

[30] Y. Li et al., *Directed Evolution of Human T-Cell Receptors with Picomolar Affinities by Phage Display. Nature Biotechnology*, Nat. Biotechnol., 23 (2005) 3.

[31] D.T. Harris et al., *Deep Mutational Scans as a Guide to Engineering High Affinity T Cell Receptor Interactions with Peptide-Bound Major Histocompatibility Complex*, J. Biol. Chem., 291 (2016) 47.

Barbara Bravi: fisica teorica di formazione, è attualmente Lecturer in Biomathematics presso il Dipartimento di Matematica di Imperial College London. Ha conseguito il dottorato presso King's College London nel 2016, ed in seguito ha effettuato periodi di ricerca post-dottorale a École Politechnique Fédérale de Lausanne ed a École Normale Supérieure de Paris. Si interessa di inferenza statistica, processi stocastici e reti complesse con applicazioni in biologia.