Trovare un gene responsabile del tumore

Bernard Prum

Gli sviluppi della biologia moderna e soprattutto quelli della genetica molecolare richiedono dei nuovi strumenti matematici. Un esempio è fornito dal ruolo della statistica nella ricerca di un gene legato al tumore al seno.

Innumerevoli malattie hanno una componente ereditaria: il rischio che un individuo ha di esserne colpito è più o meno elevato a seconda che egli sia portatore di un gene detto di *suscettibilità* alla malattia in questione. È per questo che la genetica di oggi cerca di comprendere il ruolo dei diversi geni con particolare attenzione alla loro importanza nella eziologia dei malati. La speranza è quella di poter mettere a punto un piano di terapia in futuro. Prendiamo come esempio il tumore al seno che, in Francia, colpisce circa una donna su otto. Accanto ad alcuni fattori di rischio (alimentazione, tabacco, esposizione alle radiazioni, etc...), si è identificato da qualche anno un gene le cui mutazioni sono rintracciate nelle donne affette da tumore al seno. Questo gene è stato chiamato BRCA1 (per "breast cancer 1"). Tale risultato, di natura biomedica, non può essere ottenuto che grazie ad una serie di analisi statistiche che, come vedremo, hanno permesso di localizzare il gene in modo sempre più preciso. La genetica ha a lungo ignorato la materia

In questa mammografia, con i colori alterati, è visibile, in rosa, un tumore cancerogeno. Una parte della ricerca sul cancro al seno è rivolta al suo aspetto genetico. La statistica gioca qui un ruolo decisivo. (Negativo Kiugs College School/SPL/Cosmos)

di cui sono fatti i geni. È da poco più di vent'anni che si ha accesso in modo massiccio alle sequenze di DNA, la catena di molecole che realizza l'informazione genetica trasmessa dai genitori ai figli. Tuttavia, l'ignoranza della composizione chimica dei geni non ha affatto impedito di ottenere risultati raffinati sull'eredità di certi caratteri.

La prima domanda che ci si pone di fronte ad una malattia come il tumore (nel caso che stiamo esaminando, quello al seno) è: "È una malattia genetica? Esistono, cioè, dei geni che provocano una predisposizione a questa malattia?". Per il tumore la risposta è stata a lungo incerta. Ci si aspetta una risposta positiva se si riscontrano alcuni casi della stessa malattia in una stessa famiglia e si può quindi attribuire alla figlia o alla sorella di una donna colpita da una malattia un rischio più alto di contrarla rispetto alla media della popolazione. Per molto tempo lo statistico genetico ha avuto come dati di base alberi genetici come quello di figura 1.

Che fare di un tale albero genetico? Si è detto, appena dopo Mendel, che un carattere ereditario è spesso determinato da un "gene" in grado di prendere diverse forme, chiamato *allele*. Ogni individuo eredita un allele da suo padre ed uno dalla madre mediante una combinazione genetica che avviene nel concepimento. Quindi il genetista propone un modello di trasmissione della malattia che suppone l'intervento di certi geni ed alleli. Il compito dello statistico è di valutare questo modello con l'aiuto di documenti appropriati che permettono, ad esempio, di

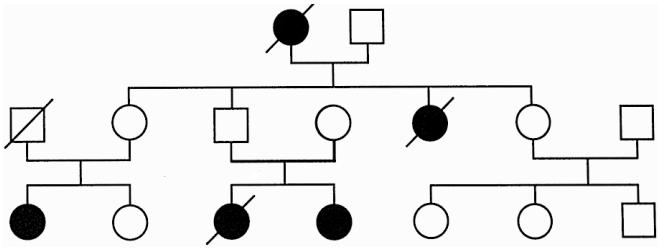

Figura 1. Una famiglia in cui si osserva una concentrazione di cancro al seno. I quadrati indicano gli uomini, i cerchi le donne. Un individuo è indicato in nero se ne è malato, in barrato se defunto. Si vede che la nonna, una delle figlie e tre nipoti hanno avuto un cancro. Beninteso, la malattia può presentarsi ad altri membri della famiglia. È a partire da tali studi che i genetisti sono stati portati a supporre l'esistenza di geni suscettibili alla malattia.

eliminare le ipotesi più semplici come quella che la malattia non abbia alcuna componente genetica. Dai casi di malattie più studiate si passa a quelle di eziologia complessa (come i tumori al seno), dove intervengono dei fattori ambientali la cui incidenza dipende dall'età del soggetto: è opportuno trattare i dati in dipendenza dal tempo e bisogna dunque fare appello alla statistica dei processi. È un ramo elaborato della matematica che si basa in gran parte sui risultati di teoria delle probabilità ottenuti dalla scuola francese negli anni '80 (P. A. Meyer, J. Jacod) e quelli di statistica, principalmente dovuti alla scuola scandinava.

Le statistiche per determinare il cromosoma portatore del gene

Una volta stabilita, dall'analisi dell'albero genetico, l'esistenza di un gene di

suscettibilità per il tumore al seno, la seconda tappa consiste nel localizzarlo, almeno grossolanamente, in uno dei 23 cromosomi presenti nell'uomo. Per questo si hanno a disposizione dagli anni '80 dei marcatori: essi sono delle piccole catene di DNA ben determinate che si possono leggere a minor "costo", grazie ad una analisi chimica piuttosto rapida. Essendo relativamente facili da localizzare, i marcatori permettono per esempio di valutare la somiglianza fra zone di cromosomi esaminate in persone malate e consanguinee. Più grande è la somiglianza di una stessa zona di cromosoma presente in persone consanguinee infette, più elevata è la probabilità che questa zona contenga un gene implicato nella malattia.

Ma una tale analisi, puramente statistica, è complicata dal fatto che ogni genitore non trasmette ai figli i cromosomi che egli ha ereditato dai genitori, ma una "ricombinazione" di questi (Figura 2). Se consideriamo due geni situati all'inizio su uno stesso cromosoma, essi potrebbero, dopo la ricombinazione, trovarsi su due cromosomi diversi. La probabilità che questo accada è in proporzione più alta che se i due geni fossero lontani. Analizzare il tasso di uguaglianza in un cromosoma è quindi un processo aleatorio. Grazie alla statistica dei processi si può individuare un intervallo lungo il quale è possibile trovare un gene di suscettibilità. L'uso dei marcatori ha così permesso al gruppo americano di Jeff M. Hall, a Berkeley, di localizzare nel 1990 il gene BRCA1 sul cromosoma 17.

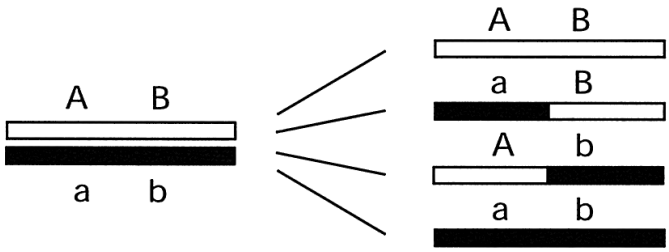

Figura 2. Per ogni coppia di cromosomi di un individuo, un cromosoma è ereditato da suo padre (in nero nella figura) e l'altro da sua madre (in bianco). Un genitore trasmette ad ogni figlio un solo cromosoma di ciascuna specie. Ma, prima della trasmissione, i cromosomi di ogni coppia possono scambiarsi "pezzi" in modo casuale. Questo processo, detto di ricombinazione fa sì che ogni genitore trasmetta al figlio un cromosoma ricombinato (in una delle quattro possibilità indicate nella figura, in cui si suppone che i cromosomi si scambino due pezzi).

Leggere la molecola di DNA per decifrare completamente il gene e le sue forme anormali

Poi si tratta di localizzare in modo più preciso il gene e determinarne la struttura. Si dice che il DNA, il materiale che contiene l'informazione genetica, è una lunga catena molecolare "scritta" in un alfabeto di 4 lettere (A,C,G,T), iniziali di quattro tipi di molecole con cui esso è formato. Le banche genetiche riportano molti miliardi di tali lettere (non ne arrivano che 25 milioni al giorno...).

La precisione del metodo dei marcatori permette in media di localizzare un gene su una sequenza di DNA contenente all'incirca 4 milioni di lettere. Per sapere esattamente quale allele o quale mutazione è responsabile, per esempio, del

tumore al seno, bisogna "leggere" queste sequenze su soggetti sani e confrontarle con quelle lette su individui malati.

Questo permette di trovare un "errore di battuta" in un testo di 4 milioni di caratteri, una quantità paragonabile a quanti ce ne sono in un libro di 2000 pagine o, meglio in tanti libri di 2000 pagine quanti sono gli individui da studiare. Questo procedimento è oneroso perfino con mezzi informatici potenti. Ora nell'uomo i geni non costituiscono più del 3% dei cromosomi. Il resto del materiale cromosomico è qualificato come *intergenico*. Se si riuscisse a limitare la ricerca degli errori di battitura ai soli geni, si ridurrebbe la sequenza da esplorare ad una trentina di pagine, il cui confronto è cosa agevole per un qualsiasi calcolatore.

Ma come è possibile distinguere i geni dal resto? Si dimostra che lo "stile" con cui sono scritti i geni è diverso dallo "stile" intergenico: le frequenze di successione delle lettere non sono le stesse. Si può cercare di sfruttare questa differenza di stile per *annotare* la frequenza e distinguere i geni dalla parte intergenica. La sfida è ardua. Si deve fare appello a dei modelli statistici chiamati *catene di Markov nascoste*, sviluppati negli anni '80, in rapporto specialmente con dei problemi di riconoscimento automatico delle parole; essi sono stati adattati alla genomica e sono stati messi a punto, nello stesso tempo, degli algoritmi capaci di caratterizzare le differenze di "stili" e di attribuire uno stile ad ogni posizione sul cromosoma.

È così che si è riusciti a localizzare precisamente il BRCA1. Oggi si può anzi leggerlo facilmente su ogni malato. Questo gene di suscettibilità al tumore al seno è composto da 5592 lettere e se ne conoscono più di 80 alleli. Resta un nuovo lavoro per lo statistico: stabilire le relazioni fra i diversi alleli e l'incidenza di questo tumore.

La biologia offre alla matematica un nuovo terreno di azione

L'esempio del gene BRCA1 suggerisce che, nei prossimi anni, la biologia giocherà nei confronti della matematica un ruolo paragonabile a quello svolto dalla fisica nel corso di buona parte del XX secolo: offrirà un campo di applicazione ai recenti progressi tecnici e provocherà l'elaborazione di nuovi strumenti (qui abbiamo riportato solo gli sviluppi statistici, ma avremmo potuto riportare altri campi della matematica come i sistemi dinamici, l'ottimizzazione e la geometria; infatti la conformazione spaziale delle molecole gioca un ruolo essenziale nella loro funzione). Una nuova sfida oggi è stata lanciata allo statistico: attualmente è possibile disporre qualche migliaia di reagenti su una superficie di vetro di un centimetro quadrato (detti "pulci") e sapere così quali geni agiscono ed in quali tessuti in quelle date condizioni sperimentali o ...in quelle delle cellule tumorali. Le misure effettuate in laboratorio, in centinaia di condizioni diverse, forniscono ai ricercatori un numero considerevole di dati numerici che

caratterizzano l'espressione di migliaia di geni. Ad oggi, solo delle analisi statistiche possono pretendere di elaborare una così grande mole di dati sperimentali. Questo potrebbe portare a precisare i legami tra geni e malattie.

Bernard Prum
Laboratorio Statistico
e Genomico (UMR CNRS 8071).
La Génopole, Università di Évry.

Alcuni riferimenti bibliografici:

- B. Prum, "*Statistique et génétique*" *dans Development of Mathematics* 1950-2000 (sous la dir. de J.-P. Pier, Birkhäuser, 2000).
- C. Bonaïti-Pellié, F. Doyon et M. G. Lé, "*Où en est l'épidémiologie du cancer en l'an 2001*", Médecine-Science, 17, pp. 586-595 (2001).
- F. Muri-Majoube et B. Prum, "*Une approche statistique de l'analyse des génomes*", Gazette des mathématiciens, n. 89, pp. 63-98 (juillet 2001).
- B. Prum, "*La recherche automatique des gènes*", La Recherche, n. 346, pp. 84-87 (2001).
- M. S. Waterman, *Introduction to computational biology* (Chapman & Hall, 1995).