Eur. Phys. J. B **14**, 783–786 (2000)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Classificazione automatica delle conformazioni delle anse degli anticorpi

A.M. Lesk[a]

University of Cambridge Clinical School, Wellcome Trust Centre for the Study of Molecular Mechanisms in Disease, Cambridge Institute for Medical Research, Wellcome/MRC Building, Hills Road, Cambridge CB2 2XY, UK

Ricevuto il 10 Agosto 1999

Sommario. I siti di legame con l'antigene degli anticorpi sono costituiti principalmente da sei anse, le RDC (Regioni determinante la complementarietà). La conformazione della catena principale di cinque di queste anse è descritta dal modello delle "strutture canoniche", secondo il quale le catene principali di queste anse possiedono un repertorio di conformazioni limitato e discreto. Le conformazioni osservate per l'ultima ansa (RDC3 della catena pesante) sono molto più variabili, e qui ci limitiamo a distinguere fra le conformazioni di questa regione più vicine all'impalcatura. Dato il rapido aumento nella popolazione di strutture delle immunoglobuline nella banca dati, sarebbe utile avere un metodo automatico per identificare le strutture canoniche di un anticorpo. Abbiamo scritto un programma che accetta in entrata le coordinate di una struttura di un anticorpo, e assegna le strutture canoniche delle anse. Il programma utilizza (1) un insieme delle strutture di anticorpi conosciute e la definizione delle RDC in tutte queste strutture, e (2) le coordinate degli esempi che ci forniscono la definizione di ciascuna struttura canonica.

Abstract. The antigen-binding sites of antibodies are constructed principally from six loops, the CDRs (complementarity-determining regions). For five of these loops the main-chain conformations follow the "canonical structure" model, according to which the mainchains of these loops have a small, discrete and limited repertoire of conformations. The observed conformations of the last loop (CDR3 of the heavy chain) are much more variable, and here we only distinguish among conformations of the portions of this region adjacent to the framework. Given the rapid growth in population of antibody structures in the protein data bank, it would be useful to have an automatic method for identifying the canonical structures of a new structure. We have written a program that takes as input the structure of an antibody and reports the assignment of canonical structures to its loops. The program uses (1) a set of known antibody structures and the definitions of the CDRs in all these structures, and (2) coordinates of examples that define each canonical structure of each loop.

PACS. 87. Biological and medical physics – 87.14.Ee Proteins – 87.15.Cc Folding and sequence analysis

1 Introduzione

Gli anticorpi rappresentano una famiglia di proteine di estremo interesse tanto per il loro ruolo nella difesa contro diverse malattie, quanto per la loro utilità nell'ingegneria proteica – per esempio, nella costruzione di enzimi artificiali (anticorpi catalitici).

Un anticorpo tipico, o immunoglobulina G (IgG), contiene quattro catene – due identiche catene leggere (L), formate da due domini chiamati V_L e C_L, e due identiche catene pesanti (H = "Heavy" in inglese) formate da quattro domini chiamati V_H, C_{H1}, C_{H2} e C_{H3} (Fig. 1). Si possono identificare due tipi di catene leggere: κ e λ. Nell'uomo, ambedue i tipi appaiono in proporzioni simili; nel topo, predomina il tipo κ.

Alle estremità N-terminali delle catene leggere e pesanti si trovano i domini variabili, V_L e V_H, che legano l'antigene. Questi domini contengono una struttura formata da due foglietti β, impacchettati l'uno contro l'altro, chiamata impalcatura. Il sito di legame con l'antigene è costituito principalmente da sei anse che collegano tra di loro i fili dei foglietti β dell'impalcatura: tre nel domino V_L – chiamate RDC1, RDC2, e RDC3 della catena leggera (oppure L1, L2, L3), e tre nel domino V_H – chiamate RDC1, RDC2, e RDC3 della catena pesante (oppure H1, H2, H3), dove RDC sta per "Regione Determinante la Complementarietà". Queste relazioni strutturali sono illustrate in Figura 1.

In che modo le sequenze aminoacidiche degli anticorpi determinano la struttura del sito di legame con l'antigene, e, di conseguenza, la specificità? La conformazione della catena principale di cinque di queste

[a] e-mail: aml2@mrc-lmb.cam.ac.uk

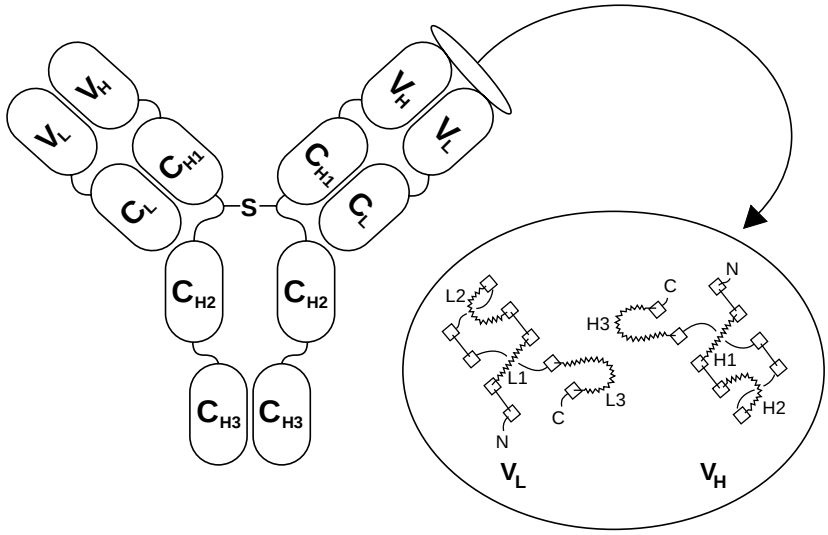

Fig. 1. Struttura di una immunoglobulina G e sito di legame con antigene. A sinistra: Le quattro catene che costituiscono le immunoglobuline di tipo G: due catene leggere, formate dai domini V_L e C_L, e due catene pesante, formate dai domini V_H, C_{H1}, C_{H2} e C_{H3}. S indica uno ponte disolfuro fra le catene pesante. A destra: La disposizione spaziale delle regioni determinanti la complementarietà.

[Structure of an immunoglobulin G and the antigen-binding site. Left: the four chains that make up the immunoglobulin of type G: two light chains, formed of domains V_L and C_L, and two heavy chains, formed of domains V_H, C_{H1}, C_{H2} and C_{H3}. S indicates a disulphide bridge. Right: the spatial disposition of the complementarity-determining regions.]

sei anse è descritta dal modello delle "strutture canoniche" definito da Chothia and Lesk [1]. Secondo questo modello, le catene principali di ciascuna di queste anse possiedono un repertorio di conformazioni limitato e discreto. Le anse assumono la loro conformazione, ciascuna dal proprio repertorio, in base alla presenza nella loro sequenza di aminoacidi di pochi residui particolari. Questi residui formano legame idrogeno, o interazioni idrofobiche (impacchettamento), oppure possono adottare conformazioni eccezionali della catena principale. (Glicina o asparagina, per cui è permessa una conformazione con $\phi > 0, \psi > 0$, al di fuori di quelle comunemente osservate nel grafico "Sasisekharan-Ramachandran," oppure prolina, che può più facilmente assumere un legame peptidico *cis*.) Le lunghezze delle anse, nella maggior parte dei casi, sono fra 3 e 10 residui ma, di solito, la scelta della struttura canonica è determinata da due residui soltanto.

Il fatto che solo pochi residui determinano la conformazione dalla catena principale di queste anse è molto interessante, in quanto permette agli altri residui delle regioni ipervariabili di variare liberamente, determinando così la topografia e la distribuzione di carica della superficie del sito di legame con l'antigene.

Il modello delle strutture canoniche fornisce una esauriente descrizione delle conformazioni della catena principale per le anse L1, L2, L3, H1 ed H2. L'ultima ansa, H3, si comporta invece in modo alquanto diverso, come risultato del meccanismo attraverso il quale i geni degli anticorpi sono formati.

Nell'uomo ed in molti altri animali, la regione del gene per la catena pesante, che codifica per il domino V_H, è costruita dalla combinazione dei segmenti genetici V, D e J. Ciascun gene per il domino V_H è il risultato di tre segmenti V, D e J, ciascun scelto fra parecchie possibilità che appaiono nel DNA; inoltre, una giunzione imprecisa aumenta la diversità di questa regione. H3 corrisponde alla regione nella sequenza genetica dell'anticorpo dove i segmenti V, D e J si riuniscono. (In qualche caso, il segmento D codifica esattamente per il segmento H3 dell'anticorpo.)

Di conseguenza, le conformazioni osservate per le anse H3 sono molto più variabili di quelle delle altre anse, ed il problema di classificare tali conformazioni è quindi più difficile ed elusivo. Qui ci limitiamo a distinguere fra le conformazioni della regione dell'ansa H3 prossimale all'impalcatura, chiamata da Morea *et al.* "il torso" [2].

Quando il modello delle strutture canoniche era stato inizialmente proposto, non era possibile sapere esattamente fino a che punto le conformazioni delle anse con la stessa struttura canonica sono conservate fra immunoglobuline diverse. All'epoca, le banche dati non erano ricche come adesso; contenevano solo poche strutture, molte delle quali determinate a bassa risoluzione, oppure contenenti errori abbastanza seri. Al giorno d'oggi, più di duecento strutture di anticorpi sono state determinate, e nel recente lavoro Al-Lazikani, Chothia and Lesk [3] hanno dimostrato, servendosi solo di strutture determinate ad alta risoluzione e ben raffinate, che gli esempi della stessa struttura canonica delle anse L1, L2, L3, H1 ed H2 di immunoglobuline diverse hanno la stessa conformazione della catena principale, le differenze tra queste (misurate dalla radice della deviazione quadrata media delle

posizioni degli atomi della catena principale dopo una sovrapposizione ottimale) essendo inferiori ad 1.0 Å.

2 Un programma per l'identificazione automatica delle strutture canoniche

Dato il rapido aumento della popolazione di strutture nella banca dati, sarebbe utile avere un metodo semplice ed automatico per identificare le strutture canoniche delle anse di una nuova struttura di una immunoglobulina. Abbiamo quindi scritto un programma in grado di fare questo, e l'abbiamo provato sulle strutture conosciute delle immunoglobuline presenti nella Banca dati Delle Proteine (PDB) [4]. È opportuno sottolineare che questo lavoro non consiste nella *predizione* della conformazione di queste strutture; il nostro scopo è molto più semplice: *data* la struttura sperimentale, ne analizziamo automaticamente la conformazione delle anse.

Il programma accetta in entrata le coordinate di una struttura di un anticorpo, e assegna le strutture canoniche delle anse. Il programma ha bisogno (1) di un insieme delle strutture di anticorpi conosciute e della definizione delle RDC in tutte queste strutture, e (2) delle coordinate degli esempi che ci forniscono la definizione di ciascuna struttura canonica ("olotipi") (v. Tab. 1).

Questo programma si occupa di due problemi: il primo consiste nell'identificazione, nella nuova struttura, delle regioni che corrispondono alle anse determinanti la complementarietà; il secondo consiste nell'identificazione della struttura canonica per ciascuna ansa.

Per estrarre le RDC dalla struttura dell'anticorpo, è necessario allineare i suoi residui con quelli di altri anticorpi per i quali sono già particolareggiate le regioni delle RDC. Questo può essere fatto in due modi: o (1) allineando *la sequenza* aminoacidica del nuovo anticorpo con quelle di anticorpi le cui strutture sono già note, o (2) allineando *le strutture* corrispondenti. Il metodo dell'allineamento strutturale è più robusto perchè le strutture delle impalcature dei domini variabili degli anticorpi sono molto simili; le loro sequenze, invece, pur essendo anch'esse molto ben conservate, possono cambiare – perdendo, in qualche caso, anche i residui più conservati che in generale utilizzeremmo come punto di riferimento per generare l'allineamento delle sequenze: per esempio, le cisteine.

Il secondo problema è più facile da affrontare. Abbiamo una libreria di strutture canoniche – un insieme di strutture per ciascuna ansa – e, dato l'allineamento di sequenze, si può effettuare la sovrapposizione ottimale delle anse del nuovo anticorpo su tutte le strutture canoniche con la stessa lunghezza che quest'ansa può assumere. È noto che il valore della radice della deviazione quadrata media della catena principale fra esempi diversi di una struttura canonica è di solito inferiore a 1.0 Å; quindi, se il minimo valore della deviazione fra un'ansa del nuovo anticorpo e una delle strutture canoniche per quest'ansa è inferiore a questo valore, si può assegnare questa struttura canonica a tale ansa.

Tabella 1. (a) Le strutture canoniche della catena leggera. (b) Le strutture canoniche della catena pesante.

[(a) The canonical structures of the light chain. (b) The canonical structures of the heavy chain.]

(a)

Ansa	tipo del domino	struttura canonica	lunghezza	esempio	risoluzione Å
L1	κ	1	6	2fbj	1.95
L1	κ	2A	7	1fvc	2.2
L1	κ	2B	7	1vfa	1.8
L1	κ	3	13	1hil	2.0
L1	κ	4	12	1flr	1.85
L1	κ	5	11	1ggi	2.8
L1	κ	6	8	1fig	3.0
L1	λ	1	10	2rhe	1.6
L1	λ	2	11	7fab	2.0
L1	λ	3A	11	1ind	2.2
L1	λ	3B	11	1mfa	1.7
L1	λ	4	8	8fab	1.8
L2	$\kappa + \lambda$	1	3	2rhe	1.6
L3	κ	1	6	1flr	1.85
L3	κ	2	6	2fbj	1.95
L3	κ	3	5	1bql	2.6
L3	κ	4	4	1dfb	2.7
L3	κ	5	7	1baf	2.9
L3	κ	6	5	1eap	2.5
L3	λ	1A	6	1ind	2.2
L3	λ	1B	6	7fab	2.0
L3	λ	1C	6	1mfa	1.7
L3	λ	2	8	2rhe	1.6

(b)

Ansa	tipo del domino	struttura canonica	lunghezza	esempio	risoluzione Å
H1	γ	1	7	1mfa	1.7
H1	γ	2	8	1baf	2.9
H1	γ	3	9	1ggi	2.8
H2	γ	1	3	1vfa	1.8
H2	γ	2A	4	1mfa	1.7
H2	γ	2B	4	2cgr	2.2
H2	γ	3A	4	8fab	1.8
H2	γ	3B	4	1ind	2.2
H2	γ	3C	4	1igm	2.3
H2	γ	4	6	1flr	1.85
H3	γ	TB	> 10	1mfa	1.7
H3	γ	TNB-1	> 10	1flr	1.85
H3	γ	TNB-2	> 10	1mrf	2.4
H3	γ	corto	≤ 10	1mrf	2.4

3 Descrizione e controlli del programma

Il nostro programma per l'assegnamento automatico delle strutture canoniche delle anse di una nuova struttura

anticorpale si serve di due librerie di strutture: (1) quella dei domini di immunoglobuline, e (2) quella delle strutture canoniche. Ci serviamo della libreria di strutture per generare gli allineamenti strutturali, così da poter estrarre le anse desiderate, e della libreria di strutture canoniche per assegnare le strutture canoniche alle anse identificate nel passaggio precedente.

Inizialmente, il programma esegue l'allineamento strutturale fra la nuova immunoglobulina e ciascuna struttura nella libreria di domini (il metodo è descritto nel riferimento [5], ed in altri lavori ivi citati) e ne sceglie due: uno fra i domini V_L ed un altro fra i domini V_H, con i valori minimi della radice della deviazione quadrata media con i rispettivi domini della nuova immunoglobulina. Il risultato è una lista dei residui corrispondenti fra la nuova struttura ed i domini noti estratti dalla libreria. Dato che vogliamo identificare le regioni ipervariabili, e che queste regioni spesso *non* s'allineano con i domini conosciuti, il programma identifica i residui della nuova struttura che corrispondono a quelli dell'impalcatura della struttura nota più vicini alle anse, e definisce come RDC tutti i residui compresi tra quelli della nuova struttura. Troviamo interessante che sia difficile, solo dagli allineamenti strutturali dei domini V_L, distinguere fra domini V_L di tipo κ e λ. Per farlo bisogna controllare le conformazioni delle anse. (Sarebbe facile distinguere fra κ e λ dalle sequenze del domino CL – il domino C-terminale della catena leggera – ma spesso non disponiamo della sequenza C_L, per esempio, nei casi in cui sia stata determinata solamente la struttura di un frammento F_v, il quale contiene soltanto i domini V_L e V_H.)

Il programma effettua quindi la sovrapposizione ottimale fra le anse della nuova struttura e tutte le strutture canoniche proprie di questa ansa con la stessa lunghezza (metodo descritto in [6]), e riporta i valori della radice della deviazione quadrata media delle posizioni degli atomi della catena principale. Per ciascuna delle anse ipervariabili nella nuova struttura, il programma identifica la struttura canonica nota per questa RDC, con la stessa lunghezza e con il valore minimo della radice della deviazione quadrata media. Se questo valore è inferiore ad 1.0 Å, il programma assegna la struttura canonica dell'ansa della libreria all'ansa della nuova struttura. Se questo valore minimo è superiore ad $1.5 - 2.0$ Å, si è probabilmente scoperta una nuova struttura canonica (ma questo non è accaduto nel corso di questo lavoro). Valori minimi fra 1.0 Å ed 1.5 Å (anch'essi non ottenuti nel corso di questo lavoro) indicherebbero probabilimente una conformazione di un'ansa imparentata ad una struttura canonica nota, ma distorta. Bisognerebbe indagare tali casi individualmente, che ci aspettiamo comunque essere piuttosto rari.

Nel primo passagio, per la libreria dei domini noti, abbiamo osservato che basta servirsi dei domini di solo tre strutture conosciute: D1.3 (1vfa) e J539 (2fbj) (le cui catene leggere sono di tipo κ), e KOL (2fb4) (catena leggera di tipo λ).

Abbiamo provato il programma utilizzando le strutture (ad eccezione delle tre che costituiscono la libreria dei domini noti) alle anse delle quali Al-Lazikani *et al.* [3] hanno assegnato le strutture canoniche (Immunoglobulina [Codice PDB]): RHE (2rhe), NEWM (7fab), SE155 (1mfa), CHA255 (1ind), HC19 (1gig), HIL (8fab), POT (1igm), H52 (1fgv), 4D5 (1fvc), 17/9 (1hil), McPC603 (2imm), 4–4–20 (1flr), TE33 (1tet), NC8.6 (2cgr).

In ciascun caso, il programma riproduce correttamente l'assegnamento delle strutture canoniche alle anse L1, L2, L3, H1 ed H2 come pubblicato da Al-Lazikani *et al.* [3] e classifica correttamente la conformazione del "torso" di H3 [2]. È notevole che il programma può trattare tre casi che presentano difficoltà particolari: 2rhe e 2imm, che contengano soltanto catene leggere, ed 7fab, in cui manca la regione L2.

4 Conclusione

Un programma per l'assegnamento automatico delle strutture canoniche alle anse di una nuova struttura di un anticorpo dimostra che le regole che definiscono le strutture canoniche si esprimono con abbastanza chiarezza da potere essere utilizzate come base di un procedimento oggettivo e di applicazione generale. Inoltre, questo programma è utile per analizzare nuove strutture di immunoglobuline. Il metodo può essere applicato a problemi diversi, trattandosi del riconoscimento dei motivi strutturali nelle famiglie di strutture proteiche.

Ringrazio la Dott.ssa V. Morea per l'aiuto nella preparazione del manoscritto, e The Wellcome Trust per i fondi.

Bibliografia

1. C. Chothia, A.M. Lesk, J. Mol. Biol. **196**, 901–917 (1997).
2. V. Morea, A. Tramontano, M. Rustici, C. Chothia, A.M. Lesk, J. Mol. Biol. **275**, 269–294 (1998).
3. B. Al-Lazikani, A.M. Lesk, C. Chothia, J. Mol. Biol. **273**, 927–948 (1997).
4. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, J. Mol. Biol. **112**, 535–542 (1977).
5. A.M. Lesk, *Integrated Access to Sequence and Structural Data. In Biosequences: Perspectives and User Services in Europe*, edited by C. Saccone (EEC, Bruxelles 1986), pp. 23–28.
6. M. Rustici, A.M. Lesk, J. Comp. Biol. **1**, 121–132 (1994).