



HAL
open science

Séquençage de novo de la région variable (FV) des anticorps circulants : objectif d'un logiciel de réassemblage

Bertrand Saunier, Nicolas Maillet

► To cite this version:

Bertrand Saunier, Nicolas Maillet. Séquençage de novo de la région variable (FV) des anticorps circulants : objectif d'un logiciel de réassemblage. Bulletin de l'Académie Vétérinaire de France, 2023, 176, 10.3406/bavf.2023.71035 . pasteur-04345253

HAL Id: pasteur-04345253

<https://pasteur.hal.science/pasteur-04345253>

Submitted on 14 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

SÉQUENÇAGE *DE NOVO* DE LA RÉGION VARIABLE (Fv) DES ANTICORPS CIRCULANTS : OBJECTIF D'UN LOGICIEL DE RÉASSEMBLAGE

DE NOVO SEQUENCING OF THE VARIABLE REGION (FV) OF CIRCULATING ANTIBODIES: AIM OF REASSEMBLY SOFTWARE

Par Bertrand SAUNIER¹ et Nicolas MAILLET²

(Communication présentée le 25 mai 2023, manuscrit accepté le 18 mai 2023)

RÉSUMÉ

L'immunité humorale est essentielle à la réponse adaptative aux infections et à leur prévention par la vaccination. La diversité des cellules productrices d'anticorps a été bien mieux étudiée que celle des anticorps proprement-dits, alors qu'ils constituent la pièce maîtresse de la réponse humorale. Or, un décalage important a été récemment observé entre le répertoire des récepteurs antigéniques des lymphocytes B circulants et celui des anticorps du sérum. En permettant le séquençage *de novo* d'anticorps, la protéomique en *bottom-up* présente de nombreux atouts pour combler cette lacune. Les clivages protéiques utilisés par cette méthode augmentent nettement la couverture de séquençage. Ceux-ci impliquent néanmoins de reconstruire *a posteriori* la séquence des chaînes lourdes et légères des immunoglobulines et de les rapparier. Cette procédure devient rapidement complexe avec des mélanges d'anticorps, comme ceux issus du sérum. C'est pourquoi nous développons un logiciel pour évaluer le taux maximum de séquençage d'anticorps atteignable par l'approche *bottom-up*.

Mots-Clés : Immunité humorale ; anticorps sériques ; spectrométrie de masse ; séquençage *de novo* ; reconstruction de séquences ; « One Health ».

ABSTRACT

Humoral immunity is key for the adaptive response to infections and their prevention by vaccination. Although antibodies are at the core of the humoral response, their diversity has been much less studied than that of antibody-producing B cells. But a discrepancy has recently been reported between the repertoire of antigen receptors carried by circulating B cells and that of serum antibodies these cells produce. By enabling the *de novo* sequencing of antibodies, *bottom-up* proteomics has many assets to fill this gap. The protein cleavages that define this method markedly increase sequencing coverage. However, this requires reconstructing *a posteriori* the sequences of the heavy and light chains of the immunoglobulins and re-matching them. The complexity of this procedure increases considerably when analyzing complex antibody mixtures, such as those contained in serum. Therefore, we are developing software to assess the maximum antibody sequencing rate achievable with the *bottom-up* approach.

Keywords: Humoral immunity; serum antibodies; mass spectrometry; *de novo* sequencing; sequence reconstruction; « One Health ».

(1) M.D., Ph.D., ORCID 0000-0002-0700-3082, Unité de Virologie Structurale, Institut Pasteur, 28 rue du Dr Roux, 75015 Paris,
Courriel : bsaunier@pasteur.fr.

(2) Ph.D., ORCID 0000-0003-1611-5243, Hub de Bioinformatique et Biostatistiques, Institut Pasteur, 28 rue du Dr Roux, 75015 Paris,
Courriel : nicolas.maillet@pasteur.fr.

INTRODUCTION

Au XVIII^e siècle, la variole était une maladie fort répandue qui tuait un tiers de ceux qui la contractaient. Développer cette maladie protégeait les survivants contre une réinfection. Une telle « immunité » justifia le recours à l'inoculation d'extraits de pustule variolique, une pratique nommée « variolisation ». Malheureusement, il arrivait fréquemment que les personnes y étant soumises décèdent. D'autres observations suggéraient que des paysans contaminés par la vaccine, ou variole de la vache (vache = *vacca* en latin), non seulement ne développaient pas d'infection sévère, mais étaient de surcroît protégés de la variole humaine. L'apport d'Edward Jenner fut de démontrer l'efficacité de cette « vaccination » qui était plus sûre que la variolisation, et d'en répandre la pratique.

Au siècle suivant, Louis Pasteur inventa l'atténuation des agents infectieux par passages successifs sur des supports biologiques permissifs sous-optimaux. Il utilisa ce principe qui complétait celui inventé par Jenner, pour développer son « vaccin » contre la rage. Pasteur l'utilisa pour la première fois en 1885 sur Joseph Meister qui avait été mordu par un chien enragé mais qui, grâce au vaccin, ne développa pas cette maladie mortelle. En 1901, le premier Prix Nobel de Physiologie ou Médecine fut attribué à Emil Adolf von Behring pour sa découverte d'un sérum antitoxine de la diphtérie et l'étude rationnelle des maladies. L'ère de l'immunologie moderne était née.

Ainsi, Hans Buchner (1890), von Behring (1890) et Paul Ehrlich (1892 et 1900) mirent en évidence l'importance de l'immunité humorale dans la réponse aux maladies infectieuses, notamment à travers le concept d'interaction anticorps-antigène. Ils montrèrent que les anticorps étaient le principal support de l'immunité adaptative. Celle-ci est articulée avec l'immunité innée qu'elle complète, et cible plus spécifiquement des antigènes portés par les agents pathogènes. En 1948, Astrid Fagraeus découvrit que des lymphocytes B - principalement les plasmocytes - sécrétaient les anticorps.

Depuis cette époque, la réponse humorale a été étudiée :

- De façon indirecte, avec des fractions de sérum provenant de sujets immunisés, mais dont la composition en anticorps n'était pas connue ;
- De façon plus spécifique, avec des immunoglobulines (Igs) produites par des cellules B isolées *in vitro*, mais sans véritablement connaître le rapport entre les immunoglobulines identifiées et les anticorps circulant dans le sérum.

Rétablir le lien entre les deux a longtemps été considéré impossible. En pratique, le répertoire identifié à partir de cellules B spléniques ou circulantes a souvent été assimilé à celui circulant effectivement sous forme d'anticorps sériques. Or, depuis quelques années, les arguments s'accumulent en faveur d'un décalage, non seulement quantitatif, mais aussi qualitatif, entre ces deux répertoires (Snapkov *et al.* 2022 ; Neumeier *et al.* 2022 ; de Graaf *et al.* 2022 ; Lavinder *et al.* 2014). Ceci a conduit à envisager de nouvelles approches pour étudier la réponse humorale prenant effectivement place chez les patients, notamment au cours des maladies infectieuses.

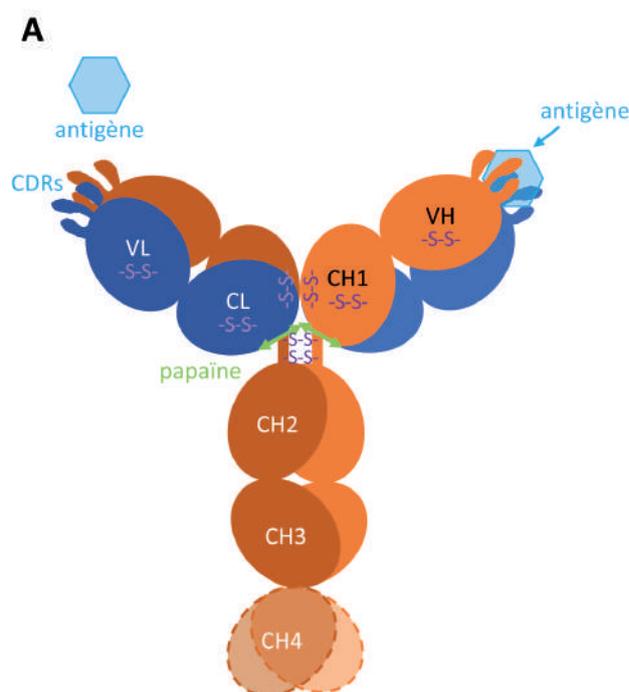
Dans cet article, nous passons brièvement en revue certains fondements de l'immunité humorale pour proposer une façon plus directe de l'étudier. Nous avons commencé à développer

un outil bioinformatique permettant d'évaluer le taux maximum d'anticorps pouvant être séquencés par cette méthode. Dans cette première partie, nous envisageons les perspectives que l'obtention de notre outil ouvrirait pour de futures études cliniques de terrain et la démarche « One Health ».

BREF APERÇU SUR LES ANTICORPS

Composition des immunoglobulines

Gerald Edelman a partagé le Prix Nobel de Physiologie ou Médecine avec Rodney Robert Porter en 1972 pour la découverte de la structure des molécules d'anticorps. Les immunoglobulines sont des protéines globulaires glycosylées composées d'hétérodimères de chaînes légères (2 domaines globulaires) et lourdes (4-5 domaines) associées entre elles par des ponts disulfures. Elles adoptent une forme en Y d'environ 10 nm (Figure 1A). Chacune des trois branches est composée de deux brins reliés par un pont disulfure vers le centre du Y. Chaque brin comporte deux (ou trois) domaines globulaires d'à peu près 110 acides aminés chacun, soit environ 440 (ou 660) acides aminés par branche. Après digestion d'une immunoglobuline par de la papaïne, les trois branches sont libérées sous forme de deux Fab identiques et d'un Fc (Figure 1B). Chaque Fab est l'hétérodimère d'une chaîne légère et de la moitié N-terminale d'une chaîne lourde ; à l'extrémité libre du Y se trouve son domaine variable, appelé région Fv, responsable de la liaison à l'antigène. Fc est un dimère de la partie restante, glycosylée, des chaînes lourdes qui, selon la classe ou l'isotype, porte la fonction effectrice de l'immunoglobuline : soit par fixation du complément (complexe C1q), soit après liaison à l'un des récepteurs de Fc (c.à.d. d'immunoglobulines) à la surface de cellules immunitaires spécialisées.



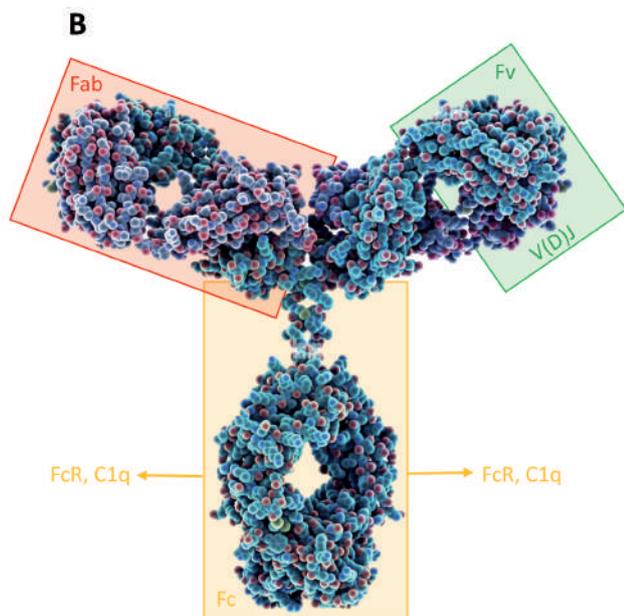


Figure 1 : Principaux domaines d'une immunoglobuline. A) Domaines globulaires des immunoglobulines : L = chaîne légère (en bleu foncé) et H = chaîne lourde (en ocre) ; VH ou VL = régions variables avec leurs boucles CDRs (cf. le glossaire) interagissant avec l'antigène, en bleu clair, et CH1, 2 et 3 (et parfois CH4) ou CL = régions constantes ; en mauve, les ponts disulfures (-S-S-) et, en vert clair, sites de coupure par la papaine. B) Domaines fonctionnels des immunoglobulines (superposés à la structure 3D d'une IgG) : en rouge, le domaine Fab = (VL-CL)-S-S-(VH-CH1) ; en jaune, le domaine Fc = (CH2-CH3)-S-S-(CH2-CH3) qui s'articule à la fonction effectrice des anticorps ; en vert foncé, la partie variable Fv = VL + VH (par recombinaison V(D)J) ; cf. le glossaire), dont on cherche à déterminer la séquence en acides aminés.

Chez l'être humain, il existe deux types de chaînes légères : kappa (κ) ou lambda (λ), et cinq types de chaînes lourdes : gamma (γ , isotypes 1, 2, 3 ou 4), delta (δ), epsilon (ϵ), alpha (α , isotypes 1 ou 2) et mu (μ). Le nombre d'hétérodimères, le plus souvent eux-mêmes reliés entre eux par un pont disulfure, varie selon le type d'immunoglobulines. Il y en a, respectivement, 2 pour les IgG, IgD et IgE, formant donc des tétramères de masse molaire allant de 150 à 200 kDa (1 Dalton = 1 g/mole) ; 2 ou 4 pour les IgA, selon qu'elles sont circulantes (≈ 160 kDa) ou sécrétées avec une chaîne J et un composant sécrétoire (≈ 385 kDa) ; et 10 pour les IgM, soit un pentamère de tétramères, également reliés par une chaîne J (≈ 1 MDa). Les IgG, IgD, IgE et IgA circulantes sont divalentes pour l'antigène, tandis que les IgA sécrétées et les IgM ont, respectivement, une valence de quatre et cinq, au lieu de dix pour cette dernière en raison d'un encombrement stérique. Dans le sérum, les IgG sont les plus courantes ($\approx 3/4$ des Igs).

Mécanisme de diversification des immunoglobulines

Si chaque anticorps était encodé par un gène, le système immunitaire aurait besoin de millions de gènes pour produire les différents anticorps (chez l'humain, le nombre total de

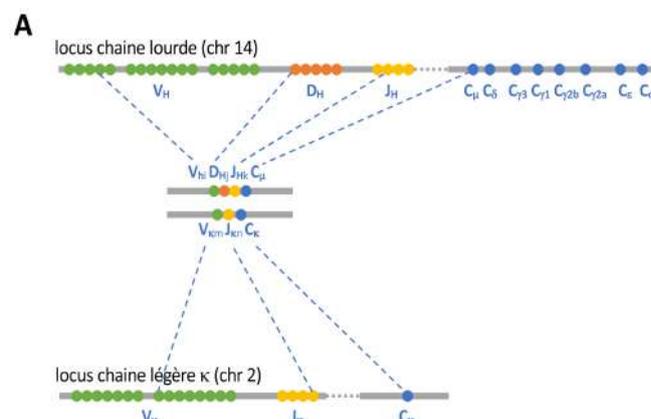
gènes est compris entre 30 et 80 mille). Susumu Tonegawa a obtenu le Prix Nobel de Physiologie ou Médecine en 1987 pour sa « découverte du principe génétique de la diversification des anticorps ». Il a montré que le matériel génétique des lymphocytes B de souris adultes est recombinaison en comparaison de celui de souris embryonnaires et que cela explique la diversité de la région variable des anticorps (Fv).

Chez l'être humain, les séquences génomiques encodant les chaînes lourdes (H = *heavy*) sont portées par une région du chromosome 14, et les chaînes légères (L = *light*) par les chromosomes 22 (pour les chaînes kappas) et 2 (pour les lambdas). La variabilité de séquence des chaînes légères est obtenue par recombinaison (cf. le glossaire) entre une séquence de région VL et une de région JL, parmi celles présentes dans le génome. Pour les chaînes lourdes, ce sont deux recombinaisons qui ont lieu entre trois régions : une VH, une DH et une JH (Figure 2A) ; ceci augmente considérablement la variabilité des chaînes lourdes par rapport aux chaînes légères.

Néanmoins, le génome de chaque individu ne comporte qu'une fraction de la diversité identifiée pour chacune de ces régions (environ 1/10 pour les régions VH) ; cette diversité est fortement influencée par la sélection naturelle en lien aux maladies infectieuses. Les jonctions se formant entre les régions V(D)J sélectionnées au moment de la recombinaison sont elles-mêmes sujettes à forte variabilité (Figure 2B). Les chaînes ainsi recombinaison font l'objet d'une variabilité additionnelle sous forme de mutations, insertions et/ou délétions touchant préférentiellement des boucles à l'extrémité libre des Fv : nommées régions hypervariables ou CDRs (*complementarity-determining regions*), ces boucles sont le principal support des interactions anticorps-antigène (Figure 1A).

Différenciation des cellules B et production d'anticorps

Les deux chaînes, lourde et légère, sélectionnées dans une même cellule B naïve définissent ce qu'on appelle une lignée germinale (Nutt *et al.* 2015). Les anticorps sont d'abord exprimés sous forme de récepteurs d'antigène à la surface des cellules B naïves (BCR = *B cell receptor*). La rencontre avec un antigène approprié - impliquant l'affinité de liaison au BCR et l'interaction avec des cellules T *helper* elles aussi spécifiques de l'antigène - déclenche l'apparition d'un signal stimulateur à l'intérieur de la cellule. Celle-ci se multiplie alors et la séquence de ses régions Fv peut être diversifiée par des hypermutations somatiques (Figure 3).



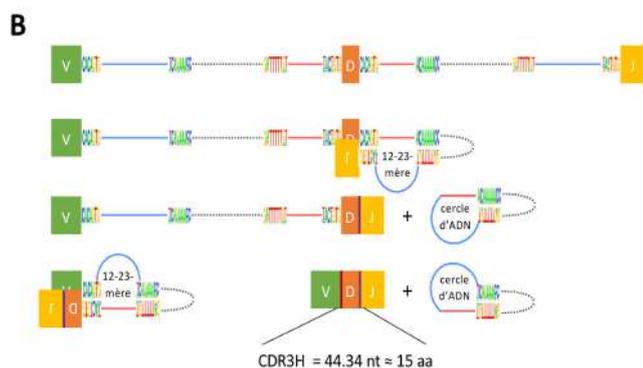


Figure 2 : Diversification des chaînes d'immunoglobulines. A) Recombinaisons V(D)J et Jc : un seul gène est sélectionné parmi chaque groupe de gènes V, (D) et J sur le chromosome (chr) correspondant pour encoder les chaînes lourdes (en haut) et légères (en bas, ici κ) d'une immunoglobuline ; le nouveau gène V(D)J est alors recombéné avec celui d'une région constante. B) Mécanisme d'une recombinaison V(D)J : après chaque recombinaison entre sites RSS donneurs et accepteurs (cf. le glossaire), la machinerie de réparation d'ADN excise le 12-23-mère (écart d'un tour d'hélice entre les segments d'ADN rouge et bleu) avec tous les gènes intermédiaires (en pointillés), puis raccorde les gènes adjacents en introduisant des nucléotides au hasard (traits verticaux, en rouge et mauve) : la longueur moyenne du CDR3 de chaîne lourde chez les humains est indiquée (nt = nucléotide ; aa = acide aminé).

Lorsque cela résulte en un réel gain d'affinité, de spécificité ou de réactivité croisée pour l'antigène, les cellules concernées prolifèrent davantage. Plusieurs cycles de maturation peuvent prendre place dans les centres germinatifs des organes lymphoïdes (nœuds lymphatiques, rate), jusqu'à atteindre une différenciation :

- Ré-activable en cas de réinfection : cellules B mémoires circulantes (de longue durée de vie) ;
- Terminale pour les cellules productrices d'anticorps : plasmoblastes circulants (de courte durée de vie) et plasmocytes nichant dans la moelle osseuse (de longue durée de vie) ; ces derniers sécrètent de bien plus grandes quantités d'anticorps.

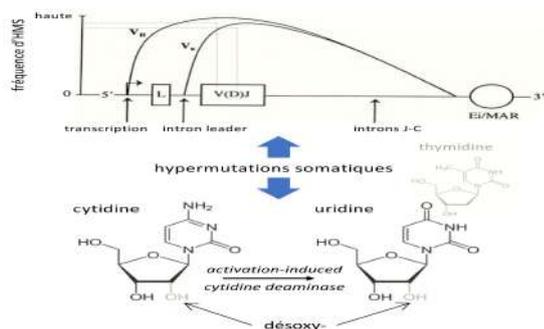


Figure 3 : Maturation des régions Fv des immunoglobulines (après recombinaison V(D)J). En haut : en abscisse, la distance sur le gène recombéné est représentée à partir du site de départ de la transcription et, en ordonnée (HMS), la fréquence des hypermutations somatiques (cf. le glossaire). En bas : la maturation des gènes d'immunoglobuline se fait par mutations de déoxycytidines en désoxyuridines (lues comme des thymidines) ; une fois ces mutations transcrits dans l'ARNm, les codons modifiés pourront, ou non (redondance du code génétique), changer des acides aminés lors de la traduction ; l'affinité de l'anticorps pour l'antigène peut s'en trouver inchangée, augmentée ou diminuée.

APPLICATION A L'ÉTUDE DE L'IMMUNITÉ HUMORALE

Identification d'anticorps monoclonaux

L'immunité humorale est essentielle à la réponse aux infections, à l'adoption d'une immunité vaccinale, et peut aussi être impliquée dans certains processus auto-immunitaires. Il a rapidement paru intéressant d'étudier les répertoires d'anticorps activés en réponse aux différents antigènes concernés. Le séquençage utilisé par Edelman était fastidieux et ne permettait pas d'étudier directement la diversité des anticorps impliqués. En revanche, la découverte de Tonegawa permettait d'envisager l'immunisation de souris (ou d'autres animaux) avec les antigènes d'intérêt, en vue de sélectionner les cellules B produisant des anticorps spécifiquement dirigés contre ces antigènes. Néanmoins, chaque antigène peut comporter plusieurs épitopes (domaine de liaison à l'anticorps) et donc être la cible d'anticorps assez variés. Il était donc nécessaire d'avoir une vision plus précise pour comprendre les mécanismes de la réponse immunitaire. L'idée directrice était notamment d'obtenir des outils de plus en plus spécifiques pour le diagnostic et la thérapeutique.

Dans les années 1970, il a été découvert un cancer de cellules B (myélome) qui produisait de grandes quantités d'anticorps identiques. Ceci conduisit au développement de la technique d'immortalisation des lymphocytes B issus de la rate, en les fusionnant avec des cellules de myélome. Les cellules filles provenant toutes d'une même cellule mère —ou clones cellulaires produisant un unique type d'immunoglobuline— sont alors sélectionnées par dilutions successives. En 1984, Georges Köhler et César Milstein ont reçu le Prix Nobel de Physiologie ou Médecine pour cette technique de production d'anticorps monoclonaux. Cette technique a essentiellement été utilisée chez la souris, en raison de son format plus adapté aux études en laboratoire. Cela a permis d'obtenir de nombreux outils diagnostiques et de recherche, mais la spécificité antigénique des anticorps monoclonaux de souris limitait leur application thérapeutique, notamment chez l'humain. Fut alors envisagée l'humanisation d'anticorps monoclonaux, c.à.d. échanger leur partie constante pour leur contrepartie humaine, tout en conservant inchangée la portion Fv spécifique de l'antigène. Ceci avait pour but de limiter la réponse allogénique contre des immunoglobulines provenant d'une autre espèce. Mais la spécificité de réponse chez les humains présente d'autres différences avec celle des souris. Par exemple, les boucles des CDRs —notamment du CDR3H— sont plus courtes chez la souris, ce qui les empêche d'accéder à certains épitopes masqués par de la glycosylation, tandis qu'ils peuvent rester accessibles aux anticorps humains.

Production d'anticorps monoclonaux recombinants

De facto, cela ouvrit la porte à la sélection d'anticorps monoclonaux, non plus à partir des cellules B de la rate, mais circulantes, donc plus facilement applicable aux humains. Avec les progrès du séquençage, obtenir la composition des transcrits de ces cellules était devenu moins fastidieux et moins cher. Il devenait alors possible de produire les anticorps de façon dite recombinante, par transfection de constructions plasmidiques appropriées dans des

lignées de cellules en culture, plus faciles à manipuler. Ceci conduit à développer des techniques plus sophistiquées pour isoler les cellules circulantes d'intérêt.

L'une des plus courantes est le marquage d'antigène par une sonde fluorescente : après incubation avec des cellules nucléées provenant d'un échantillon de sang (PBMC = *peripheral blood mononuclear cells* ; cf. le glossaire) de volume suffisant pour être représentatif (> 100 mL), celles-ci sont triées selon l'intensité de lumière qu'elles réémettent une fois exposées à une source laser de longueur d'onde susceptible d'exciter la sonde. Les cellules les plus marquées sont celles exprimant à leur surface la plus grande quantité de BCRs de haute affinité. Avec les trieurs modernes, il est possible d'ajouter des critères de sélection supplémentaires (par ex., double marquage de l'antigène ; détection de CD79a, un marqueur de BCRs). Après clonage des cellules positives (ou pluri-positives), on procède au séquençage des gènes recombinés ou de leurs produits, les ARNm encodant les chaînes lourde et légère. Leurs séquences nucléotidiques sont alors co-exprimées dans des cellules en culture (par exemple, 293T ou CHO) qui les sécrètent dans le milieu, d'où elles sont purifiées par liaison à de la protéine bactérienne A et/ou G.

Cette approche a permis le développement de nombreux outils diagnostiques, mais aussi thérapeutiques, souvent utilisés en cancérologie ou contre des maladies infectieuses (par exemple, durant la pandémie de Covid-19). Lorsque ces anticorps ont vocation à être injectés à des patients, la protéomique permet d'effectuer un contrôle de qualité. Est notamment vérifiée la séquence protéique qui pourrait se trouver changée du fait de l'apparition éventuelle de mutations, d'une délétion ou d'un *editing* de l'ARNm, mais surtout par l'existence de modifications post-traductionnelles ou d'une troncation protéolytique indésirables.

APPROCHE PROPOSÉE POUR SÉQUENCER LES ANTICORPS DU SÉRUM

Positionnement du problème

Lavinder *et al.* (2014) ont constaté un écart important entre la diversité des ARNm de chaînes d'immunoglobuline séquencés par NGS (*next-generation sequencing*), à partir de cellules B circulantes, et les quantités d'IgG détectées dans le sérum par spectrométrie de masse (MS ; Figure 4). Les IgG circulantes étaient pourtant censées dériver de la traduction de transcrits présents dans ces cellules. En fait, cette observation suggérait, non seulement que certaines cellules B circulantes ne sécrètent pas d'anticorps (ce qui n'était pas vraiment une surprise), mais aussi qu'une proportion non négligeable de ces immunoglobulines proviendrait de cellules qui ne circulent pas (un fait déjà admis pour les plasmocytes) (Snapkov *et al.* 2022).

Plus récemment, le nombre d'IgG circulant en quantité significative (Guthals *et al.* 2017 ; Lavinder *et al.* 2014) - encore appelées anticorps immuno-dominants a été estimé être bien plus faible (Bondt *et al.* 2021) que ce qui était anticipé sur la foi de la diversité des cellules B en capacité de produire des anticorps (Kitaura *et al.* 2017). Cette observation a d'ailleurs été étendue à l'ensemble des immunoglobulines sériques, dont le nombre total ne dépasserait pas quelques centaines, mais dont le profil

évoluerait très vite en fonction des besoins (de Graaf *et al.* 2022). Il existe donc un véritable besoin de séquençage direct des anticorps du sérum pour étudier la réponse immunitaire humorale de façon plus pertinente.

Plusieurs technologies ont été envisagées pour relever ce défi :

- Séquençage d'Edelman : nous avons vu qu'il était long, fastidieux et pas approprié à des mélanges complexes ;
- Nanopores : cette technologie lancée en 1995 pour le séquençage nucléotidique, n'est pas vraiment opérationnelle pour le séquençage protéique ;
- Microscopie électronique : elle reste limitée à des anticorps immuno-dominants en interaction avec des antigènes, et ne permet de déterminer que la séquence des régions en contact avec ces derniers (Antanasijevic *et al.* 2022).

Toutefois, ces technologies ne permettent pas (ou pas encore) d'envisager de séquençage de mélanges contenant une grande diversité d'immunoglobulines. C'est pourquoi plusieurs auteurs ont proposé l'utilisation de la spectrométrie de masse (Figure 4) pour déterminer leur structure primaire - c.à.d. l'ordre d'enchaînement de leurs acides aminés - notamment en raison des améliorations très significatives apportées à cette technologie durant la dernière décennie (Snapkov *et al.* 2022), et de la disponibilité d'outils d'analyse de plus en plus performants, bien qu'encore nettement insuffisants pour cette indication (Schulte *et al.* 2022 ; Chen *et al.* 2020).

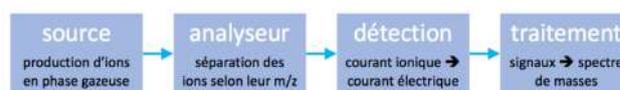


Figure 4 : Principe de fonctionnement d'un spectromètre de masse. Il réside dans la séparation en phase gazeuse de molécules chargées (ions) en fonction de leur rapport masse/charge (m/z). Selon les applications, il existe plusieurs variantes d'appareils fonctionnant selon le principe représenté.

Principe du séquençage *de novo* en spectrométrie de masse

La spectrométrie de masse est une technologie permettant de mesurer les masses isotopiques exactes (m/z) des molécules entrant dans la composition d'un mélange, avec une précision allant jusqu'à six chiffres après la virgule (détermination de la masse molaire au μg près). Cette technique est souvent utilisée pour identifier les protéines contenues dans un mélange. Après digestion d'une séquence d'acides aminés par une endoprotéase —généralement la trypsine qui coupe la liaison peptidique après une lysine (K) ou une arginine (R)— la masse exacte des peptides résultants est mesurée (Figure 5). Elles sont comparées aux masses calculées à partir des séquences protéiques connues, provenant d'une banque de données adaptées à la nature de l'échantillon : par exemple, les protéines humaines pour étudier un échantillon provenant d'un patient. Mais cette simple approche ne fonctionne plus pour des anticorps dont la séquence n'est pas connue (Muth & Renard, 2018).

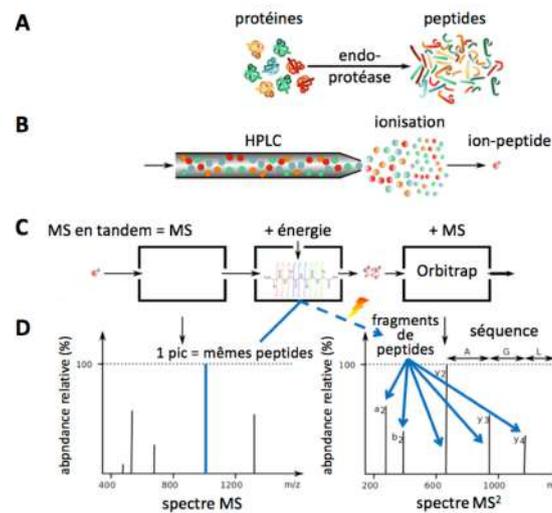


Figure 5 : Principe d'une acquisition de données en bottom-up LC-MS². A) Digestion d'un mélange de protéines par une endoprotéase (cf. le glossaire). B) Séparation des produits de digestion par HPLC et ionisation à l'entrée du spectromètre de masse (MS). C) MS en tandem (MS/MS ou MS²) : le type d'énergie utilisée conditionne les règles de fragmentation (Figure 6B). D) La première MS permet de sélectionner des ion-peptides uniques, tandis que la seconde permet de mesurer la masse exacte de leurs fragments ; la seconde mesure requiert une précision d'autant plus importante que les règles de fragmentation sont complexes (d'après Barillot et al. 2013, CC BY-SA 3.0).

Après avoir détecté un pic peptidique particulier dans le spectromètre de masse (1^{ère} mesure de m/z), il est aussi possible de l'utiliser comme source secondaire pour effectuer une nouvelle mesure. Celle-ci s'effectue après fragmentation aléatoire d'une liaison peptidique (cf. le glossaire) sur chacune des molécules du peptide composant le pic. Les règles de cassure sont dépendantes de la ou des source(s) d'énergie utili-

sée(s). Une fois celles-ci prises en compte, les différences de m/z entre fragments adjacents correspondent à la masse exacte des acides aminés rentrant dans la composition du peptide de départ. De proche en proche, on peut retrouver l'ordre des acides aminés en utilisant un logiciel adapté, une procédure appelée séquençage *de novo*, c.à.d. sans référence à une base de données (Figure 6).

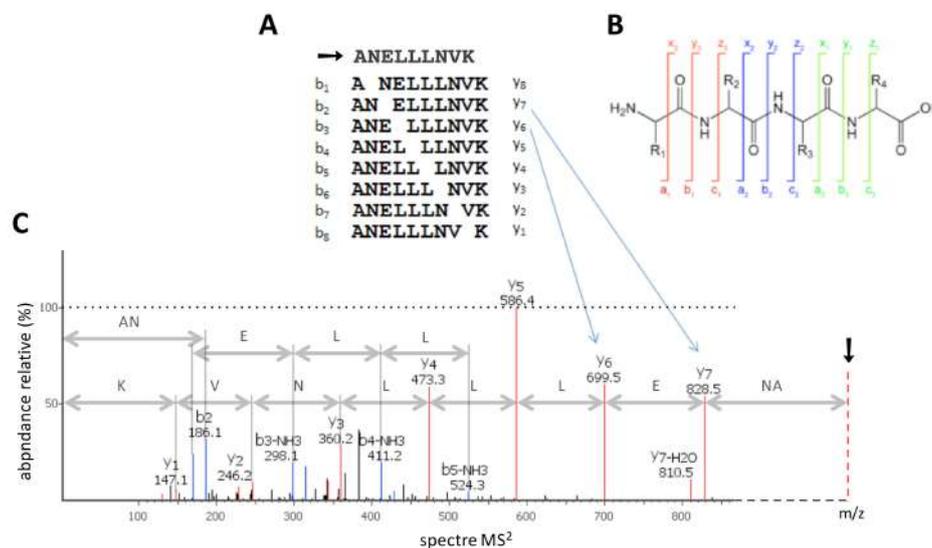


Figure 6 : Principe d'un séquençage peptidique *de novo* à partir de données en bottom-up LC-MS². A) Fragmentation aléatoire d'un ion précurseur (→) : chaque peptide-molécule est cassée une seule fois, mais pas toujours à la même position ; un pic donne donc lieu à tous les ions fils possibles, mais à des degrés divers et, en conséquence, ils peuvent ne pas être tous détectés. B) Règles de fragmentation : la position de cassure de la liaison peptidique (-CHR₁-CO-, -CO-NH- ou -NH-CHR₂-) dépend du type d'énergie utilisée (par ex., electron transfert dissociation) et conditionne les fragments peptidiques obtenus (respectivement, x/a, y/b ou z/c) ; en cas de fragmentations multiples en série, ces règles se combinent. C) Détermination de la séquence d'un ion précurseur (spectre MS) à partir du m/z de ses fragments (spectre MS²) : une fois les pics attribués aux séries, ascendante (a, b ou c) ou descendante (x, y ou z), la différence de m/z entre deux pics consécutifs correspond à celui de l'acide aminé séparant les deux fragments adjacents ; on passe ensuite aux pics consécutifs suivants, et ainsi de suite, pour obtenir au final deux séquences identiques, mais avec des directions opposées : l'ordre des acides aminés est inversé dans la série descendante (d'après He & Bin, 2010). Une fois les peptides séquencés, il faut les réassembler pour retrouver la séquence de l'Ig de départ.

Cette pratique est donc particulièrement adaptée au séquençage de protéines dont la composition exacte n'apparaît pas dans des bases de données, comme c'est le cas de la grande majorité des anticorps. Néanmoins, le séquençage *de novo* de grosses protéines ou de complexes est souvent incomplet, surtout à partir d'un mélange (Camperi *et al.* 2021 ; Cheng *et al.* 2020). C'est pourquoi on a recours à une digestion préalable de telles protéines, une approche appelée *bottom-up* LC-MS², par opposition à la précédente (sans digestion) nommée *top-down* (cf. le glossaire). En effet, le séquençage de peptides est beaucoup plus performant que celui de protéines entières (Srzentic *et al.* 2020), notamment si l'on combine différents modes de fragmentation en série (Gadush *et al.* 2022 ; Guthals *et al.* 2013 ; Frese *et al.* 2012). Toutefois, ceci complique les règles de fragmentation à prendre en compte par le logiciel d'analyse *de novo* et augmente les erreurs d'identification. C'est pourquoi, on utilise généralement un Orbitrap (Makarov, 2000), pour déconvoluer avec grande précision la masse des produits de fragmentation (cf. le glossaire). L'utilisation d'intelligence artificielle améliore notablement la qualité du séquençage *de novo* (Gueto-Tettay *et al.* 2023). Avec de telles technologies, les taux de couverture avoisinent aujourd'hui les 100%. Néanmoins, une fois la séquence des peptides de clivage obtenue, il faut reconstruire la séquence des chaînes de l'immunoglobuline.

Principe de reconstruction d'une séquence protéique

Comme déjà évoqué, le séquençage d'anticorps monoclonaux par spectrométrie de masse est de pratique relativement courante depuis plusieurs années (par ex., Creative-Biolabs). Pour les anticorps utilisés en thérapeutique chez l'humain, le régulateur américain (FDA) exige que le contrôle de qualité comporte un séquençage *de novo*. L'approche *bottom-up* nécessite de procéder à plusieurs digestions de façon à pouvoir reconstruire la séquence de la chaîne lourde et celle de la chaîne légère (Gadush *et al.* 2022). Chaque IgG à analyser est donc soumise en parallèle à plusieurs digestions de façon que les clivages résultant de chaque digestion se produisent à des positions distinctes sur la protéine.

Par exemple, étant donnée la séquence protéique simplifiée suivante (chaque lettre représente un acide aminé, réel ou fictif) : ...xxxxxxK $\alpha\beta\gamma\delta$ Dxxxxxxx..., une digestion par la trypsinine va engendrer deux peptides plus courts : ...xxxxxxK et $\alpha\beta\gamma\delta$ Dxxxxxxx..., tandis qu'une digestion par l'Asp-N qui coupe avant un acide aspartique (D) va libérer deux autres peptides : ...xxxxxxK $\alpha\beta\gamma\delta$ et Dxxxxxxx.... Après acquisition de données LC-MS² pour chacune des deux digestions, il sera possible d'établir la séquence *de novo* (sans référence à une base de données) de quatre peptides : ...xxxxxxK, $\alpha\beta\gamma\delta$ Dxxxxxxx..., ...xxxxxxK $\alpha\beta\gamma\delta$ et Dxxxxxxx.... En constatant que ...xxxxxxK $\alpha\beta\gamma\delta$ et $\alpha\beta\gamma\delta$ Dxxxxxxx... partagent une courte séquence à leurs extrémités, respectivement, COOH- et NH₂-terminales, il devient alors possible de reconstituer la séquence départ : ...xxxxxxK $\alpha\beta\gamma\delta$ + $\alpha\beta\gamma\delta$ Dxxxxxxx... = ...xxxxxxK $\alpha\beta\gamma\delta$ Dxxxxxxx..., une procédure appelée assemblage ou contiguation (cf. le glossaire).

Mais il est également envisageable d'utiliser cette approche pour des mélanges de protéines ou d'anticorps, à condition

pour ces derniers de pouvoir retrouver l'appariement de leurs chaînes. Chaque jeu d'IgG à analyser est donc soumis en parallèle à plusieurs opérations de clivages, comme expliqué ci-dessus. Les produits de chacun de ces clivages contiennent les peptides provenant des diverses immunoglobulines constituant le mélange de départ. Ces peptides sont *a priori* indiscernables et interfèrent entre eux durant la reconstruction. La complexité de reconstruction des données de séquençage issues d'une analyse en *bottom-up* augmente donc très vite avec le nombre de molécules à analyser. Nous développons donc un outil bioinformatique permettant de réassembler la séquence de chaque immunoglobuline présente dans des mélanges d'anticorps analysés en *bottom-up*.

Le développement de ce logiciel est entièrement réalisé *in silico*, en utilisant des milliers de séquences d'immunoglobulines colligées dans les bases de données. Il convient d'attirer l'attention sur le fait que cette procédure est beaucoup moins triviale que cet exposé pourrait le laisser paraître. Notamment, en raison de la création, en cours d'assemblage, de nombreuses séquences chimériques indiscernables des vraies. Cette difficulté est rarement abordée, et pourtant observée quelle que soit la technique de reconstruction utilisée. Il existe également une véritable difficulté à rapprocher les chaînes lourdes et légères. La plupart des approches proposées jusqu'à présent reviennent à effectuer un clonage de cellules circulantes, quelle que soit la technologie choisie pour en isoler le matériel génétique ou les transcrits.

APPLICATION AUX ÉTUDES CLINIQUES ?

Les techniques actuelles, faisant appel aux cellules B circulantes, demandent de larges échantillons de sang total (> 100 mL) pour les purifier à partir de PBMC (cf. le glossaire) tout en restant représentatif de la population de cellules concernées. En cas de prélèvement infectieux, selon le degré de contagiosité et la gravité de la maladie que celui-ci est susceptible de déclencher, cela peut requérir des précautions d'emploi (confinement pour prévenir le risque de contamination) ou être soumis à des restrictions réglementaires importantes, même à distance de l'épisode aigu. Par exemple, parmi les patients ayant survécu à l'épidémie de virus Ebola qui a sévi en Afrique de l'Ouest entre 2013 et 2016, certains ont conservé des réservoirs du virus qui ont pu se réactiver à distance de l'infection initiale (Garry, 2021).

Indépendamment de ces contraintes, la préparation des PBMC sur place est rarement possible (centrifugation sur gradient de Ficoll-Plaque) et une décontamination inenvisageable sans altérer les cellules. De plus, leur stockage et leur transport se font au minimum à -80° C, ce qui, comme nous avons pu le constater au début de la campagne vaccinale contre le SARS-CoV-2, pose de nombreux soucis logistiques, pas toujours surmontables. La seule chose qui serait raisonnable du point de vue infectieux serait d'extraire les ARNm sur place, mais il conviendrait alors de les décontaminer, sans savoir en quoi cela peut modifier le résultat du séquençage. De plus, leur transport s'effectue de préférence à -80° C et ceci pose de nouveau le problème du rapprochement des chaînes lourdes et légères entres-elles.

N'avoir à récolter que du sérum facilite le transport et le stockage tout en étant moins sujet à des précautions d'emploi ou à des restrictions réglementaires : le sérum se prépare sans gradient et se décontamine plus facilement (chauffage, filtration, antisept-

tiques) ; il se stocke à -20° C et se transporte à la même température. La décongélation du sérum est aussi moins problématique que celle des cellules, surtout pour une analyse en spectrométrie de masse. Du fait de la nature des échantillons analysés avec l'approche *bottom-up*, celle-ci offre une réelle possibilité de récolter, avec un minimum d'équipement, des échantillons directement sur le terrain, et ceci dès le début d'une épidémie. Pour les mêmes raisons, cette approche faciliterait le suivi longitudinal des sujets recrutés.

Le séquençage *de novo* d'anticorps polyclonaux en LC-MS² *bottom-up* constituera un apport majeur en *Discovery* (par ex., RapidNovor). Reconstruire la structure primaire de plus de la moitié des régions variables d'un mélange d'une cinquantaine d'immunoglobulines, et correctement les rapparier, permettrait déjà d'appliquer le séquençage *de novo* à la plupart des études cliniques, notamment en se limitant à une fraction d'intérêt (par ex., après liaison à un antigène viral). Une limite impor-

tante tenait jusqu'à présent à l'absence d'outil bioinformatique permettant de reconstruire suffisamment de séquences d'immunoglobulines pour espérer améliorer et adapter les protocoles d'analyse en LC-MS² *bottom-up*. C'est pourquoi nous développons un logiciel pour évaluer, après séquençage *de novo*, le taux maximum de reconstruction et de rapparierement d'anticorps atteignable par cette approche.

CONCLUSION

Il reste un grand pas à franchir avant que le séquençage *de novo* d'anticorps polyclonaux sorte des laboratoires de pointe et permette d'envisager des études de terrain (par ex., « *One Health* »). Nous espérons que le logiciel évoqué contribuera grandement à accélérer son développement. Il reste à déterminer si cela sera suffisant pour résoudre les divergences de répertoire entre les cellules B circulantes et les anticorps.

REMERCIEMENTS

Nous remercions les Drs Vét. Francis Desbrosse et Jean-Pierre Jégou pour leurs encouragements et leur soutien, ainsi que pour la relecture de ce manuscrit.

CONFLITS D'INTÉRÊTS

Aucun

COMITE D'ÉTHIQUE

Non applicable

BIBLIOGRAPHIE

- Antanasijevic A, Bowman CA, Kirchoefer RN, Cottrell CA, Ozorowski G, Upadhyay AA *et al.* From structure to sequence: Antibody discovery using cryoEM. *Sci Adv.* 2022; 8(3): eabk2039. doi : [10.1126/sciadv.abk2039](https://doi.org/10.1126/sciadv.abk2039).
- Barillot E, Calzone L, Hupé P, Vert J-P, Zinovyev A, Computational Systems Biology of Cancer. A Chapman & Hall Book: Mathematical and Computational Biology Series. CRC Press: Boca Raton, London, New York; 2013.
- Bondt A, Hoek M, Tamara S, de Graaf B, Peng W, Schulte D *et al.* Human plasma IgG1 repertoires are simple, unique, and dynamic. *Cell Syst.* 2021; 12(12): 1131-1143.e5. doi : [10.1016/j.cels.2021.08.008](https://doi.org/10.1016/j.cels.2021.08.008).
- Camperi J, Goyon A, Guillarme D, Zhang K, Stella C. Multi-dimensional LC-MS: the next generation characterization of antibody-based therapeutics by unified online bottom-up, middle-up and intact approaches. *Analyst.* 2021; 146(3): 747-769. doi : [10.1039/d0an01963a](https://doi.org/10.1039/d0an01963a).
- Chen C, Hou J, Tanner JJ, Cheng J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int J Mol Sci.* 2020; 21(8): 2873-2898. doi : [10.3390/ijms21082873](https://doi.org/10.3390/ijms21082873).
- Cheng J, Wang L, Rive CM, Holt RA, Morin GB, Chen DDY. Complementary Methods for de Novo Monoclonal Antibody Sequencing to Achieve Complete Sequence Coverage. *J Proteome Res.* 2020; 19(7): 2700-2707. doi : [10.1021/acs.jproteome.0c00223](https://doi.org/10.1021/acs.jproteome.0c00223).
- Creative-Biolabs. Disponible à : <https://www.creative-biolabs.com/next-generation-antibody-sequencing> (consulté le 04.04.2023).
- de Graaf SC, Hoek M, Tamara S, Heck AJR. A perspective toward mass spectrometry-based *de novo* sequencing of endogenous antibodies. *mAbs.* 2022; 14(1): e2079449, doi : [10.1080/19420862.2022.2079449](https://doi.org/10.1080/19420862.2022.2079449).
- Frese CK, Altelaar AFM, van den Toorn H, Nolting D, Griep-Raming J, Heck AJR *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal Chem.* 2012; 84(22): 9668-9673. doi : [10.1021/ac3025366](https://doi.org/10.1021/ac3025366).
- Gadush MV, Sautto GA, Chandrasekaran H, Bensussan A, Ross TM, Ippolito GC *et al.* Template-Assisted De Novo Sequencing of SARS-CoV-2 and Influenza Monoclonal Antibodies by Mass Spectrometry. *J Proteome Res.* 2022; 21(7): 1616-1627. doi : [10.1021/acs.jproteome.1c00913](https://doi.org/10.1021/acs.jproteome.1c00913).
- Garry RF. Ebola virus can lie low and reactivate after years in human survivors. *Nature.* 2021; 597(7877): 478-480. doi : [10.1038/d41586-021-02378-w](https://doi.org/10.1038/d41586-021-02378-w).
- Gueto-Tettay C, Tang D, Happonen



L, Heusel M, Khakzad H, Malmström J, *et al.* Multienzyme deep learning models improve peptide *de novo* sequencing by mass spectrometry proteomics. *PLoS Comput Biol.* 2023; 19(1): e1010457. doi : [10.1371/journal.pcbi.1010457](https://doi.org/10.1371/journal.pcbi.1010457)

• Guthals A, Gan Y, Murray L, Chen Y, Stinson J, Nakamura G *et al.* *De Novo* MS/MS Sequencing of Native Human Antibodies. *J Proteome Res.* 2017; 16(1): 45-54. doi : [10.1021/acs.jproteome.6b00608](https://doi.org/10.1021/acs.jproteome.6b00608).

• Guthals A, Clauser KR, Frank AM, Bandeira N. Sequencing-grade *de novo* analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptide. *J Proteome Res.* 2013; 12(6): 2846-2857. doi : [10.1021/pr400173d](https://doi.org/10.1021/pr400173d).

• He L, Bin M. ADEPTS: advanced peptide *de novo* sequencing with a pair of tandem mass spectra, *J Bioinform Comput Biol.* 2010; 8: 981-994. doi : [10.1142/S0219720010005099](https://doi.org/10.1142/S0219720010005099). Disponible à : <https://www.bioinform.com/de-novo-sequencing/> (consulté le 04.04.2023).

• Kitaura K, Yamashita H, Ayabe H, Shini T, Matsutani T, Suzuki R. Different Somatic Hypermutation Levels among Antibody Subclasses Disclosed by a New Next-Generation Sequencing-Based Antibody Repertoire Analysis.

Front Immunol. 2017; 8: 389-399. doi : [10.3389/fimmu.2017.00389](https://doi.org/10.3389/fimmu.2017.00389).

• Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI *et al.* Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci USA.* 2014; 111(6): 2259-2264. doi : [10.1073/pnas.1317793111](https://doi.org/10.1073/pnas.1317793111).

• Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Curr Opin Chem Biol.* 2015; 24: 112-120. doi : [10.1016/j.cbpa.2014.11.007](https://doi.org/10.1016/j.cbpa.2014.11.007).

• Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem.* 2000; 72(6): 1156-1162. doi : [10.1021/ac991131p](https://doi.org/10.1021/ac991131p).

• Muth T & Renard BY. Evaluating *de novo* sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform.* 2018; 19(5): 954-970. doi : [10.1093/bib/bbx033](https://doi.org/10.1093/bib/bbx033).

• Neumeier D, Yermanos A, Agrafiotis A, Csepregi L, Chowdhury T, Ehling RA *et al.* Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *Proc Natl Acad Sci*

USA. 2022; 119(18): e2113766119. doi : [10.1073/pnas.2113766119](https://doi.org/10.1073/pnas.2113766119).

• Nutt SL, Hodgkin PD, Tarlinton DM, Corcoran LM. The generation of antibody-secreting plasma cells. *Nat Rev Immunol.* 2015; 15(3): 160-171. doi : [10.1038/nri3795](https://doi.org/10.1038/nri3795).

• RapidNovor. Disponible à : <https://www.rapidnovor.com/services/antibody-protein-discovery> (consulté le 04.04.2023).

• Schulte D, Peng W, Snijder J. Template-Based Assembly of Proteomic Short Reads For *De Novo* Antibody Sequencing and Repertoire Profiling. *Anal Chem.* 2022; 94(29): 10391-10399. doi : [10.1021/acs.analchem.2c01300](https://doi.org/10.1021/acs.analchem.2c01300).

• Snapkov I, Chernigovskaya M, Sinitcyn P, Lê Quý K, Nyman TA, Greiff V. Progress and challenges in mass spectrometry-based analysis of antibody repertoires. *Trends Biotech.* 2022; 40(4): 463-481. doi : [10.1016/j.tibtech.2021.08.006](https://doi.org/10.1016/j.tibtech.2021.08.006).

• Srzentić K, Fornelli L, Tsybin YO, Loo JA, Seckler H, Agar JN *et al.* Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry. *J Am Soc Mass Spectrom.* 2020; 31(9): 1783-1802. doi : [10.1021/jasms.0c00036](https://doi.org/10.1021/jasms.0c00036).

GLOSSAIRE

Biochimie générale

• Double hélice d'ADN (alphabet de 4 désoxynucléotides : dA, T, dG et dC) → Transcription = ARNm simple brin (alphabet de 4 nucléotides (nt) : A, U, G et C) → Traduction (code génétique = 1 acide aminé [ou stop] pour 3 nt) = protéine simple brin (alphabet de 20 [ou 22] acides aminés : A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W et Y [U et O]) pouvant s'associer à d'autres protéines pour former un multimère et/ou un complexe.

• Les protéines comportent des chaînes d'acides aminés reliés entre eux par des liaisons peptidiques : NH-CHR1-CO-NH-CHR2-CO-,

où C, H, N et O sont des symboles d'atomes ; R1 et R2 représentent la chaîne latérale de deux acides aminés.

• Clivage protéique : hydrolyse spécifique d'une liaison peptidique, généralement après incubation avec un enzyme (endo-protéase) coupant après, ou avant, un acide aminé déterminé, selon la réaction suivante :

NH-CHR1-CO-NH-CHR2-CO- + H₂O → NH-CHR1-COOH + NH₂-CHR2-CO-

C'est l'opération inverse de la synthèse protéique ; elle est notamment impliquée dans la digestion.

• Une recombinaison est le fait, pour une séquence d'ADN, d'échanger une portion de double hélice située entre deux sites RSS (*recombination signal sequence*, séquence en désoxynucléotides spécifique) accepteurs, par celle située entre deux sites RSS donneurs sur une autre molécule d'ADN. Bien qu'une recombinaison puisse survenir au hasard, la présence d'une recombinase adaptée augmente considérablement la vitesse de réaction en lui conférant une beaucoup plus grande spécificité. Les recombinases spécifiques des lymphocytes sont appelées RAG1 et RAG2.

Immunologie des anticorps

- PBMC (*peripheral blood mononuclear cells*) : couramment appelés leucocytes (ou globules blancs) qui comportent les polynucléaires, les monocytes et les lymphocytes. Ces derniers se divisent en lymphocytes T et lymphocytes B, de petite taille, et en lymphocytes NK (*natural killer*), de plus grande taille. En fait, les lymphocytes se trouvent principalement dans les nœuds lymphatiques et la rate.
- Recombinaison V(D)J : les immunoglobulines sont encodées par des gènes regroupés en grappes appelées V, D (seulement pour les chaînes lourdes) ou J. Dans une première étape, des copies de gènes J et D sélectionnées au hasard se recombinent au niveau de sites nommés RSS, de façon à éliminer l'ADN portant toutes les copies de J et D intermédiaires ; les brins sont raccordés par des enzymes de réparation de l'ADN. Dans une seconde étape, les processus se répète entre le gène JD recombiné et le gène V sélectionné. Pour les chaînes légères, il y a une seule recombinaison entre deux gènes J et V. Cette opération contribue à la diversité des régions variables (Fv) des deux chaînes d'immunoglobuline, dont la séquence protéique commence avec celle encodée par un gène V et se termine avec celle d'un gène J.
- Formation d'une immunoglobuline (Figure 1A) : Une fois créées, les régions Fv s'associent par recombinaison aux gènes adéquats de régions constantes (Figure 2A) ; elles sont de nouveau diversifiées par des mutations aléatoires pour façonner deux gènes uniques (encodant les chaînes lourde et légère) par cellule, constituant une lignée germinale. Ces lignées représentent moins d'un demi-pourcent des cellules de l'organisme ($\approx 37 \times 10^{13}$) et ne peuvent donc reconnaître qu'une fraction des antigènes auxquels celui-ci est susceptible d'être exposé ($> 10^{13}$). Il existe donc un biais dans la réponse immunitaire humorale, selon les gènes V(D)J distribués à la conception, mais aussi selon la façon dont ils sont mis en place, puis sélectionnés *in utero*.
- Les cellules B naïves expriment leurs gènes d'immunoglobuline (voir ci-dessus) sous forme de récepteurs d'antigène à leur surface (BCR, *B cell receptor*). Après liaison d'un antigène au BCR d'une cellule B naïve, celle-ci prolifère et développe des mutations supplémentaires (hypermutations) sur ses gènes recombinés (régions Fv). Ceci conduit à trois types de différenciation cellulaire : cellules B mémoire, plasmoblastes et plasmocytes (Nutt *et al.* 2015). Les gènes BCR des deux derniers types vont alors produire des immunoglobulines sécrétées, encore appelées anticorps. Ceux-ci circulent dans l'organisme indépendamment des cellules qui les ont produites, mais peuvent se lier à des récepteurs sur d'autres cellules immunitaires. Normalement, seule une toute petite fraction ($1 \text{ à } 2 \times 10^9$) des lymphocytes B circule dans le sang.
- Les CDRs (*complementarity-determining regions*), ou régions hypervariables, forment trois boucles sur chacune des chaînes, dont la plus forte variabilité est déterminée par une recombinaison V(D)J (= CDR3 ; les CDR1 et 2 sont entièrement inclus dans le gène V). Après repliement des chaînes lourdes et légères pour adopter leur structure définitive (leurs CDR1 et CDR3 sont alors reliés par un pont disulfure), ces boucles se situent vers l'extrémité libre de la région Fv et sont majoritairement responsables de la liaison spécifique des anticorps aux antigènes.

Spectrométrie de masse

- Le principe de la spectrométrie de masse (MS, *mass spectrometry* ; Figure 4) réside dans la séparation en phase gazeuse de molécules chargées (ions) en fonction de leur rapport masse/charge (m/z). Un analyseur détermine précisément la valeur de m/z et permet donc de calculer la masse isotopique (exacte) des molécules qui sont ionisées en entrée. On peut articuler deux MS en tandem (MS/MS ou MS²) via une chambre de fragmentation (Figure 5C). Ce montage permet de compléter la première mesure de l'objet étudié en l'exposant à une source d'énergie qui le fragmente : l'analyse des morceaux ainsi engendrés sert à confirmer l'identification de l'objet.
- Un analyseur Orbitrap (Makarov, 2000) est une chambre sous vide bordée par une électrode cylindrique et centrée par une électrode axiale, dans laquelle les ions sont injectés par vagues et oscillent avec un mouvement hélicoïdal à une fréquence angulaire qui peut être mesurée avec grande précision. Or, cette fréquence angulaire ω suit la relation :

$$\omega = \sqrt{k/(m/z)}$$
 où k est la constante du potentiel de champ, ce qui par équivalence permet de déterminer la masse exacte de tous les objets circulant dans l'Orbitrap à un instant t :

$$m/z = k/\omega^2$$
- Des protéines peuvent être analysées par spectrométrie de masse sous forme entière (intacte) ou après découpage en peptides par des endoprotéases (Figure 5A). Ces deux approches ont donné les termes, respectivement :
 - **top-down** (descendante), où la protéine est physiquement fragmentée sous forme de peptides dans le spectromètre pour en analyser les fragments en MS₂ (voire en MS₃ pour un séquençage), et
 - **bottom-up** (ascendante), car la protéine est coupée en peptides préalablement à son analyse en MS ; une analyse en MS₂ peut venir conforter la première identification en *shotgun* (par référence à une base de données).
 En pratique, le séquençage *de novo* d'anticorps monoclonaux en *top-down* est le plus souvent partiel et celui d'anticorps polyclonaux quasi impossible (Srzentić *et al.* 2020). En routine, c'est l'approche *bottom-up* qui est utilisée pour séquencer des anticorps monoclonaux.

- HPLC (*high performance liquid chromatography*) : c'est un type de chromatographie sur colonne permettant de séparer des molécules en fonction de leur taille et de leurs propriétés physico-chimiques. La MS est souvent précédée par une HPLC lorsque l'échantillon est trop complexe, afin d'améliorer la résolution de l'analyse (LC-MS²) : des spectres MS sont alors enregistrés pendant toute la durée de l'HPLC, dont la sortie est directement reliée à l'ionisateur en entrée de la MS (Figure 5B).

Séquençage protéique

- **De novo** : se dit d'une procédure de séquençage qui ne fait pas appel à des séquences connues, donc sans recours à des bases de données (par ex., SwissProt ou GenBank).
- Principe du séquençage **de novo** en *bottom-up* LC-MS² : il est possible d'utiliser un pic peptidique détecté dans le premier analyseur (1^{ère} mesure de m/z) comme source secondaire pour effectuer une nouvelle mesure (MS²). Un pic contient de nombreux peptides identiques (1 molécule = 1 mole/6,02 x 10²³) qu'on soumet à une source d'énergie (Figure 5C) de façon qu'aléatoirement, une seule liaison peptidique par peptide soit rompue, en moyenne. Un deuxième analyseur (par ex., Orbitrap) mesure le m/z des fragments issus de l'ensemble de ces cassures (Figures 5D). Les règles de cassure sont dépendantes de la nature de la ou des source(s) d'énergie utilisée(s). Une fois prises en compte, la différence de m/z entre fragments adjacents correspond à la masse exacte d'un acide aminé (Figure 6). De proche en proche, on peut retrouver l'ordre des acides aminés de départ, en utilisant un logiciel adapté. Cette procédure se nomme séquençage *de novo*, c.à.d. sans référence à une base de données.
- Un assemblage de séquences, ou contiguation, est le fait de les fusionner par leur partie commune C, comme suit :
A-C + C-B = A-C-B.

La spécificité de reconstruction dépend notamment de la longueur et de la diversité de C.