Introduction à l'ingénierie des protéines. DEA Agroalimentaire Assurance Qualité IRAL / INRA Cycle de préformation Beyouth, octobre 2002

Thierry Chardot

DEA Agroalimentaire Assurance Qualité

IRAL / INRA

Cycle de préformation

Introduction à l'ingénierie des protéines

Thierry Chardot
Unité Mixte de Recherche INRA INA-PG
Centre de Biotechnologie Agro Industrielle
78850 Thiverval Grignon
E-mail thierry@grignon.inra.fr

Beyrouth 30 septembre 4 octobre 2002

Définitions

I Liaisons et interactions au sein de macromolécules

II Flux de l'information génétique et notions de base de la régulation de l'expression des gènes

III Biosynthèse et structure des protéines

IV Techniques générales de biologie moléculaire. Le clonage, expression homologue et hétérologue de protéines.

V La diversité Naturelle, la diversité générée

Définitions

- **Ingénierie:** Exploitation de principes scientifiques à des fins pratiques

- **Protéines:** Macromolécule organique azotée, de poids moléculaire élevé, qui donne par hydrolyse des acides aminés, et entre pour une forte proportion dans la composition des êtres vivants. (Dictionnaire Le Robert).

L'ingénierie des protéines est le résultat de l'intégration de la biologie, de la biologie structurale, et du génie génétique. L'ingénierie des protéines consiste à améliorer les propriétés de protéines existantes, à en découvrir de nouvelles. Les applications appartiendront donc aussi bien au domaine médical (découverte de nouveaux vaccins, nouveaux anticorps), qu'au domaine alimentaire (protéines végétales aux propriétés nutritionnelles améliorées), ou bien qu'aux différents domaines de l'industrie (nouvelles fibres, nouveaux catalyseurs). Il s'agit donc de modifier les propriétés de protéines d'intérêt médical, alimentaire, biotechnologique et (ou) d'obtenir ces protéines en grandes quantités, aux fins d'usage. Pour cela, il faut connaître les éléments structuraux et les méthodes de visualisation de la structure des protéines, ainsi que les méthodes d'étude de leur fonction

Les disciplines couvertes par l'ingénierie des protéines sont :
- Les techniques de la biologie moléculaire
- Les méthodes d'étude de la structure des protéines
- Les méthodes de visualisation des structures des protéines
- les méthodes d'étude de la fonction des protéines

Le spécialiste de l'ingénierie des protéines sera donc idéalement cristallographe, chimiste, biologiste, physicien et ingénieur….
Ce cours est une introduction à l'ingénierie des protéines. Il vise à rappeler les bases du transfert de l'information génétique, de la structure des protéines. L'accent sera mis sur les moyens permettant d'exprimer des protéines actives, de les modifier. On examinera finalement comment utiliser à ces fins la diversité Naturelle, ou bien la diversité générée.

I Liaisons et Interactions au sein de macromolécules

1 .1 Les liaisons covalentes

Dans le cas de molécules biologiques on notera les liaisons C-C, C-N, O-P, C-O, O-H, entre autred. Ces liaisons ont une énergie de l'ordre de (100 kcal / mole). Elles sont très stables.

1.2 La liaison hydrogène

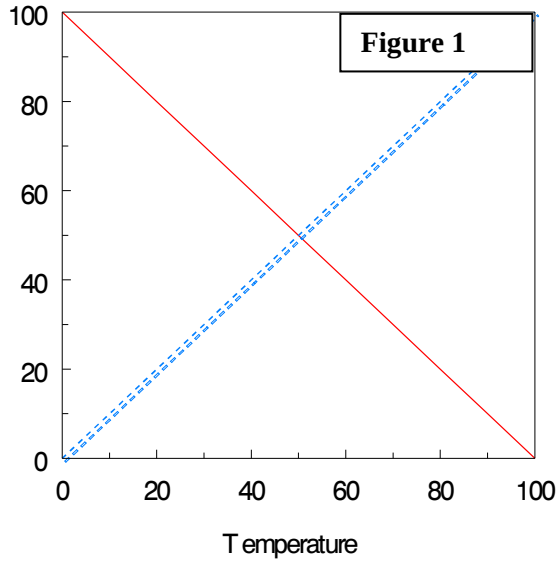
Une liaison hydrogène peut se former quand un hydrogène covalamment lié à un atome N ou O est dans le voisinage d'un autre N ou O qui possède des doublets d'électrons libres.
Ces liaisons sont peu énergétiques (-4 à –10 kcal / mole), mais très nombreuses.

1.3 Les interactions de Van der Waals

Les interactions de Van der Waals sont plus faibles que les liaisons hydrogène, portent sur des distances faibles, mais leur importance est loin d'être négligeable. Suivant la distance entre les atomes, ces forces seront attractive, ou bien répulsives.

1.4 Les interactions hydrophiles et hydrophobes

Les substances ioniques sont solubles dans l'eau car l'attraction des espèces chargée + et – pour l'eau est supérieure à l'attraction des espèces chargées de manière opposée entre elles.
Les hydrocarbures par exemple et les substances non polaires sont elles insolubles dans l'eau car les interactions eau - eau sont plus fortes que les interactions eau - substance apolaire.
Les interactions hydrophiles (trait plein) sont défavorisées par l'augmentation de la température, les interactions hydrophobes (trait pointillé) sont favorisées par l'augmentation de la température (Figure 1)



Ce sont les interactions de type non covalent qui sont responsables en très grande partie de la structure des acides nucléiques et des protéines.

II Flux de l'information génétique

Les acides nucléiques

2.1 Structure de l'ADN

L'ADN est un polymère constitué par l'enchaînement d'acides nucléiques. Les acides nucléiques au nombre de 4 sont constitués :

-D'une base azotée, d'un sucre (le désoxyribose), et d'une molécule d'acide phosphorique. Ils sont liés par des liaisons phospho diesters. La molécule d'ADN est orientée de 5' en 3. 5' correspond à l'extrémité comportant le OH du sucre qui est phosphorylé.
L'ADN peut être modifié (méthylation, restriction, …) par différentes enzymes:
C'est Watson et Crick qui on mis en évidence structure en double hélice de l'ADN. Cette hélice comporte un grand sillon et un petit sillon. Elle peut adopter plusieurs conformations. Les effets stabilisants la molécule d'ADN peuvent être classés en 4 familles :

1) L'effet hydrophobe

Les bases puriques et pyrimidiques sont orientées vers l'intérieur de la molécule. Conséquemment, les substituants hydrophiles des bases sont situés à l'extérieur et interagissent par des liaisons H avec l'eau.

2) L'empilement des paires de bases génère des interactions de Van der Waals.

Les bases sont empilées les unes sur les autres, le long de la molécule d'ADN. Elle interagissent entre elles. Les interactions sont d'intensité faible, mais elles très sont nombreuses.

3) Les liaisons H

Ce sont elles qui sont responsables de la spécificité des appariements des bases. AT GC.

4) Les interactions ioniques

Le squelette de l'ADN est chargé négativement. Il est stabilisé par des interaction avec des ions chargés + (Mg^{2+} par exemple).

On peut déplier la molécule d'ADN soit à l'aide d'agents chimiques, soit à l'aide de facteurs physiques. Si l'on chauffe de l'ADN, on va rompre des liaisons H. La température de fusion T_m est atteinte quand 50 % des appariements sont rompus. Si l'on re diminue la température de l'ADN, on reconstituera les appariements. Cette dénaturation est réversible. Attention, cela n'est

généralement pas le cas pour les protéines ! Comme on le voit sur la figure 2 un polymère constitué essentiellement de A et de T sera dénaturé à une température plus basse qu'un ADN constitué de GC. Le contenu en G+C d'un ADN donne une indication de sa température de fusion.

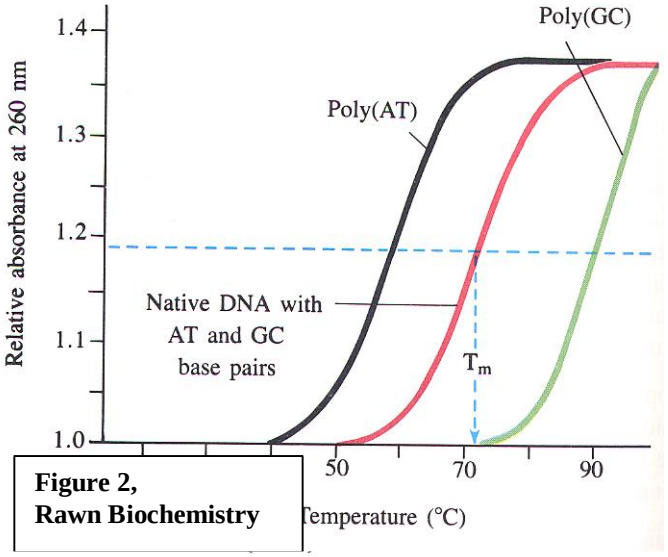


Figure 2, Rawn Biochemistry

L'ADN trouvé sous forme « nue » chez les micro organismes, est empaqueté par des protéines chez des organismes supérieurs. L'ADN peut être linéaire, ou bien circulaire. C'est le cas des plasmides. Ce sont des fragments d'ADN extra chromosomique, qui se transmettent d'un microorganisme à un autre. On les connaît à l'état sauvage. Ils ont la propriété de pouvoir transporter des fragments d'ADN qui leur sont étrangers. Ils pourront ensuite les échanger avec d'autres plasmides. Ce sont des outils de choix que nous utiliserons dans le cadre de l'ingénierie des protéines. Ils existent sous différentes formes : linéaire, enroulée, super enroulée. On peut distinguer ces différentes formes après électrophorèse en gel d'agarose (Figure 3). On les visualise alors à l'aide d'une molécule qui à la propriété de s'intercaler entre les base de la molécule d'ADN : il s'agit du bromure d'éthidium. L'ADN super enroulé semble avoir une taille inférieure à l'ADN linéarisé.L'ADN complexé au bromure d'éthidium fluoresce quand il est observé sur une table lumineuse qui émet de la lumière ultraviolette. Ce même ADN a été observé en microscopie de force atomique, et l'on voit figure 4 les différentes « formes » que peut adopter un plasmide. (Parel *et al.* British Pharmaceutical conference 2001, Abstract Book 201)

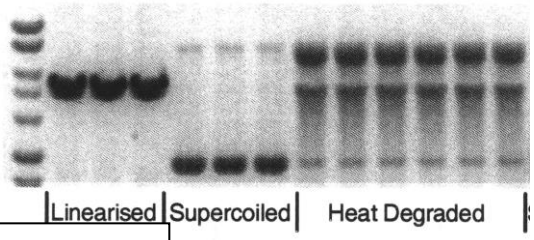
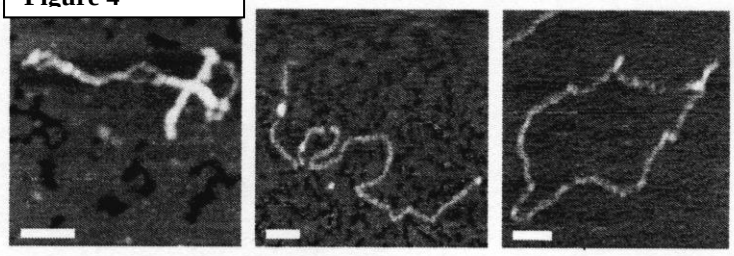


Figure 3

Figure 4



On peut aussi caractériser la pureté de l'ADN par son spectre d'absorbance. Celui est typique, car il montre un pic d'absorbance à 260 nm. Les protéines, elles absorbent la lumière vers 280

nm. La rapport d'absorbance 260/280 nm est un critère de pureté des acides nucléique. Un ADN pur aura ainsi un rapport A260/A280 > 1,8.
L'ADN est bien sûr le support de l'information génétique.

2.2 Réplication de l'ADN

La réplication permet de multiplier à l'identique les molécules d'ADN. La réplication est semi conservative. Chaque molécule issue de la réplication de l'ADN est constituée d'une molécule néo synthétisée, et d'une molécule parentale.
La réplication est bidirectionnelle. Elle se déroule au niveau de fourches de réplications, de 5' vers 3'. Sur l'un des brins, la synthèse d'ADN sera réalisée de manière continue. Sur l'autre brin, elle sera réalisée de manière discontinue. La réplication est catalysée par la ADN polymérase ADN dépendante qui a pour substrats la molécule d'ADN, les dNTP, et comme cofacteur le Mg^{2+}. Si la structure de l'ADN a immédiatement fait penser Watson et Crick à la possibilité de la transmission de l'information génétique, la structure seule de l'ADN ne pouvait révéler le contenu de l'information génétique. La question de l'utilisation de l'information restait ouverte. Il fallait faire la relation entre protéines et matériel génétique.

2.3 Structure de l'ARN.

L'ARN est un polymère constitué par l'enchaînement d'acides ribonucléiques.
Les différentes majeures entre ADN et ARN sont les suivantes
- L'hydrogène 2 du ribose de l'ADN est remplacé par un OH chez l'ARN. La présence de cet OH est responsable de la faible stabilité des ARN à pH alcalin.
- La base thymine est remplacée par un uracile dans l'ARN.
- Généralement, l'ARN est mono caténaire, mais il peut comporter des régions repliées

La molécule d'ARN est orientée de 5' en 3'. Elle comporte des régions qui vont permettre à la traduction de se dérouler correctement, telles que les séquences de Shine et Delgarno en 5', et la queue poly A, signal de terminaison de la traduction.

2.4 Transcription

La transcription est le phénomène par lequel l'information de l'ADN va être transférée de l'ADN vers les protéines avec l'ARN pour intermédiaire. L'enzyme responsable de la synthèse des molécules d'ARN est l'ARN polymérase ADN dépendante. Celle ci utilise de l'ADN comme matrice, des ribonucléotides triphosphate comme substrats et du Mg comme cofacteur. Le

$$\text{AdNn} + \text{dNTP} \xrightarrow[Mg^{2+}]{\text{polymérase}} \text{ADN n+1} + \text{PPi}$$

$$\text{PPi} + \text{H2O} \xrightarrow{\text{pyrophosphatase}} \text{P} + \text{P}$$

Figure 5

pyrophosphate (PP) est aussi un de produits de la réaction. C'est son hydrolyse qui permet l'avancement de la réaction (Figure 5 ci dessus).

Chez les organismes procaryotes, les ARN messagers peuvent coder pour plusieurs protéines : ils sont souvent poly cistroniques. Chez les eucaryotes, les ARNm codent pour une protéine : ils sont mono cistroniques. Ils peuvent contenir des séquences qui ne seront pas traduites, des introns. Ils seront alors épissés, de manière à donner un messager mature qui pourra être traduit en une séquence protéique.
L'ARN polymérase ADN dépendante reconnaît des sites de natures différentes sur la molécule d'ADN matrice. Les plus connues sont les séquences –35, la TATA box

2.5 Notions de base de la régulation de l'expression des gènes

Cas de l'opéron lactose

Les bactéries peuvent utiliser le glucose. Cependant, si on leur donne du lactose (2 glucoses) comme source de carbone, alors elle poussent moins vite. Elles synthétisent alors de la béta galactosidase. L'activité de cette enzyme n'est pas détectable quand les bactéries poussent sur du glucose (Figure 6A) . Il y a donc induction de la synthèse de cette enzyme en présence de lactose. Le système est en fait plus complexe : quand les bactéries poussent sur milieu lactose, il y a aussi induction de la synthèse de bêta galactosidase, de la lactose perméase, et d'une transacétylase.

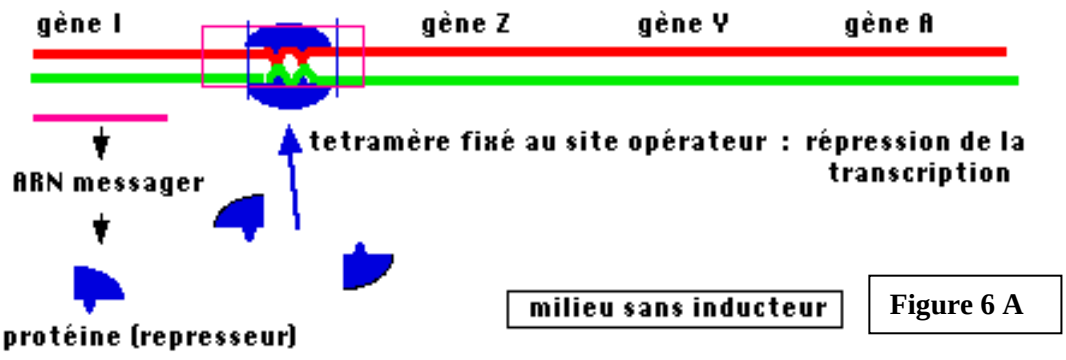


Figure 6 A

Quand la cellule est dans un milieu contenant du glucose, le répresseur est fixé sur la région opérateur de l'opéron lactose (Figure 6A), bloquant la transcription. En présence de lactose, le répresseur change de conformation, et se dissocie de la région opérateur (Figure 6B). La polymérase peut alors s'accrocher à l'ADN. La transcription peut commencer. Les différentes protéines sont alors synthétisées, la bactérie peut donc utiliser le lactose comme source de carbone. Nous reverrons plus tard dans ce cours l'intérêt d'un tel système de régulation pour l'expression de protéines recombinantes. Un des points très important à garder en mémoire, en plus du fait qu'il s'agisse d'un des premiers exemples de mécanismes de régulation de l'expression des gènes qui ait été décrit, est le suivant : ici on fait appel au fait que des protéines changent de conformation, et que ce changement de conformation est relié à un changement de d'affinité de la protéine pour la région opérateur de l'ADN.

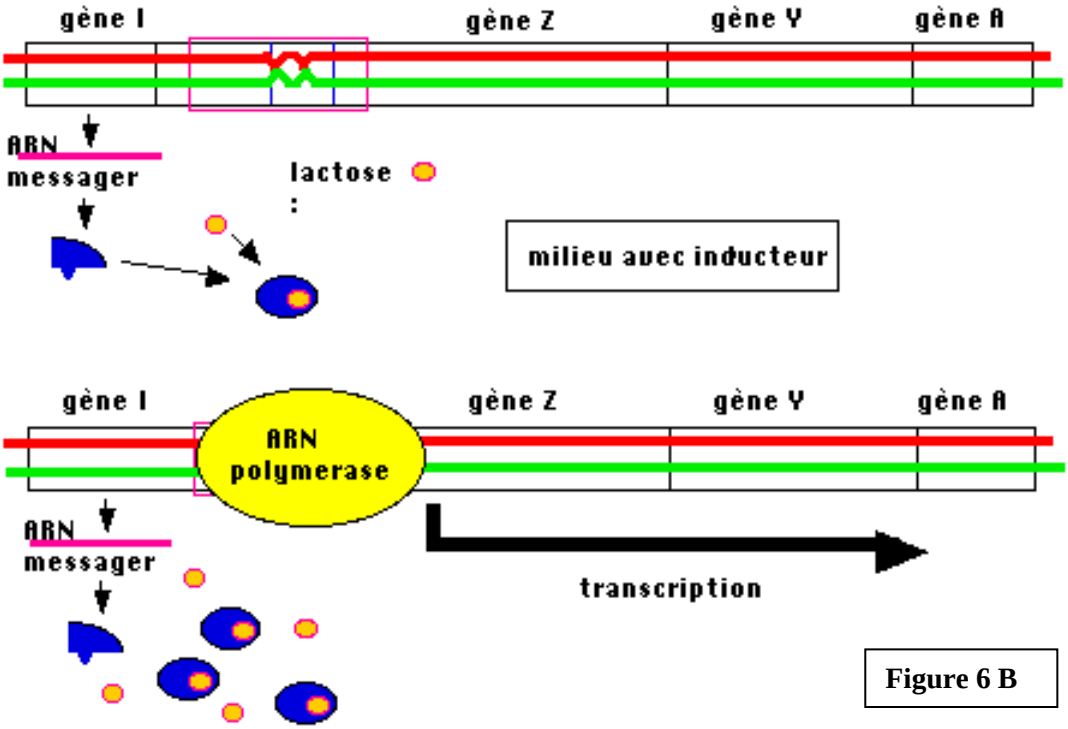


Figure 6 B

Cas des gluténines de bas poids moléculaire

Les gluténines de bas poids moléculaire (LMWG), et plus généralement les gluténines sont des protéines impliquées dans la formation du réseau de gluten de la pâte à pain. Ce réseau confère à la pâte à des propriétés viscoélastiques, et lui permet de retenir les gaz et de gonfler lors de la fermentation, de manière à donner un pain à la mie moelleuse et aérée.

Cependant, le but du blé n'est pas bien sûr de faire du pain…. Les gluténines sont des protéines de réserve. Leur dégradation lors de la germination permettra d'alimenter la plantule lors des premières étapes de son développement.

Dans la figure 7 on peut avoir une idée de la complexité d'une partie des régions promotrices d'une LMWG.

On y trouve, de 5' en 3'

La boite albumen **(I)**, qui contrôle l'expression spécifique dans ce tissu de ces protéines, est suivie d'une séquence comportant un motif homologue à l'activateur GCN4 de la levure **(II)**. On trouve alors une boîte activatrice CAAT **(III)**, puis la boîte TATAA **(IV)** qui est dupliquée, et sert de site d'attachement à la RNA polymérase ADN dépendante. On a ensuite le site d'initiation de la synthèse protéique. (**V**).

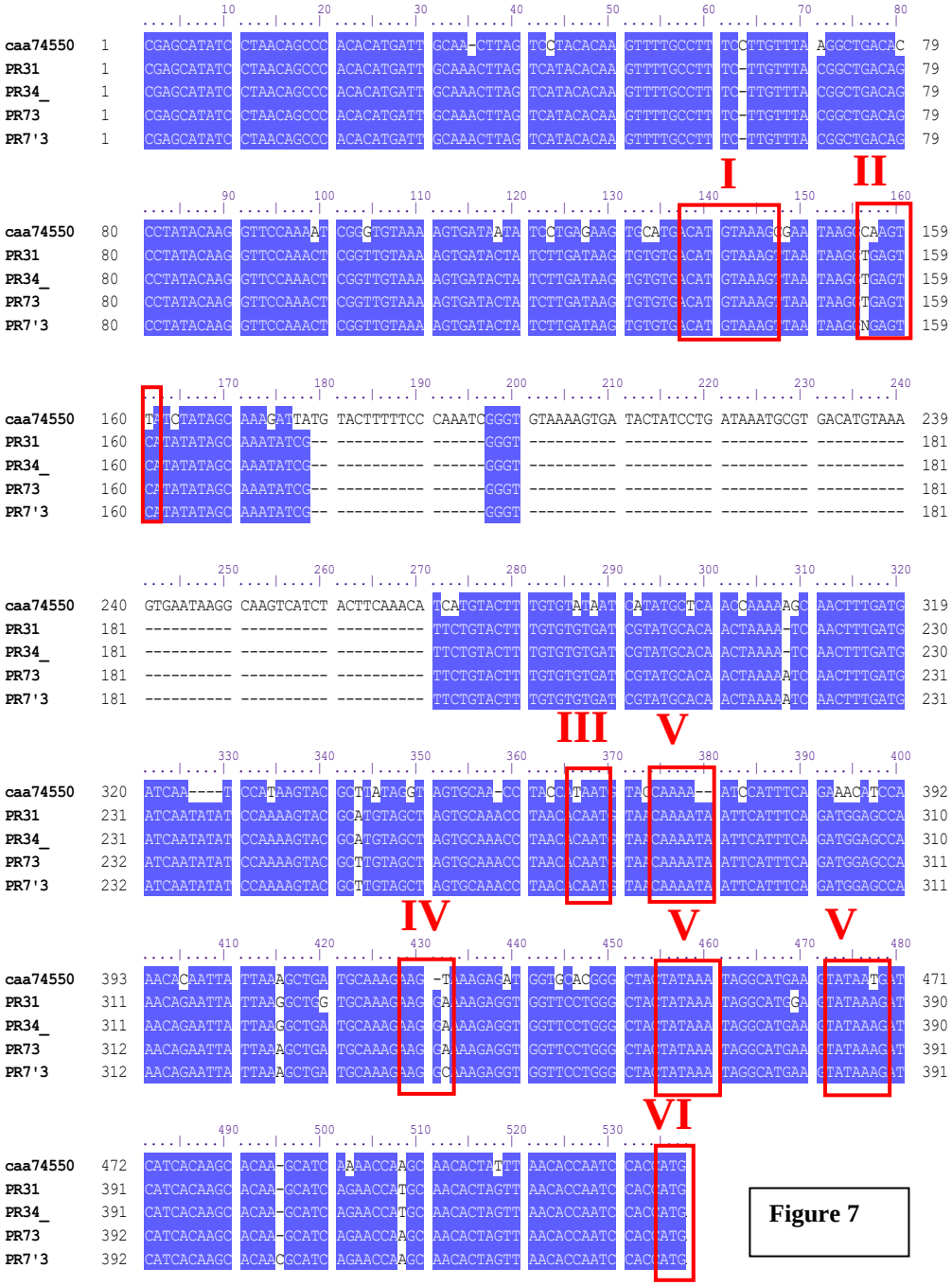


Figure 7

La structure des gluténines de faible masse moléculaire (LMWG) est très conservée. Ainsi qu'on le voit dans la figure 8, ces protéines dont on a aligné les séquences de 6 clones ont une architecture commune : Elles ont toutes une séquence N terminale qui contient le motif IPGLER, puis des zones consistant en la répétition du motif QPQQP (ou des motifs proches) pour la zone répétée I. On trouve ensuite une zone centrale, suivie d'une deuxième zone répétée consistant

en la répétition du motif QQPQQP (ou des motifs proches). Finalement, la protéine est constituée d'une zone C terminale. Les étoiles montrent la position de cystéines conservées qui permettraient de fabriquer des polymères, constitués de monomères reliés entre eux par des ponts disulfures

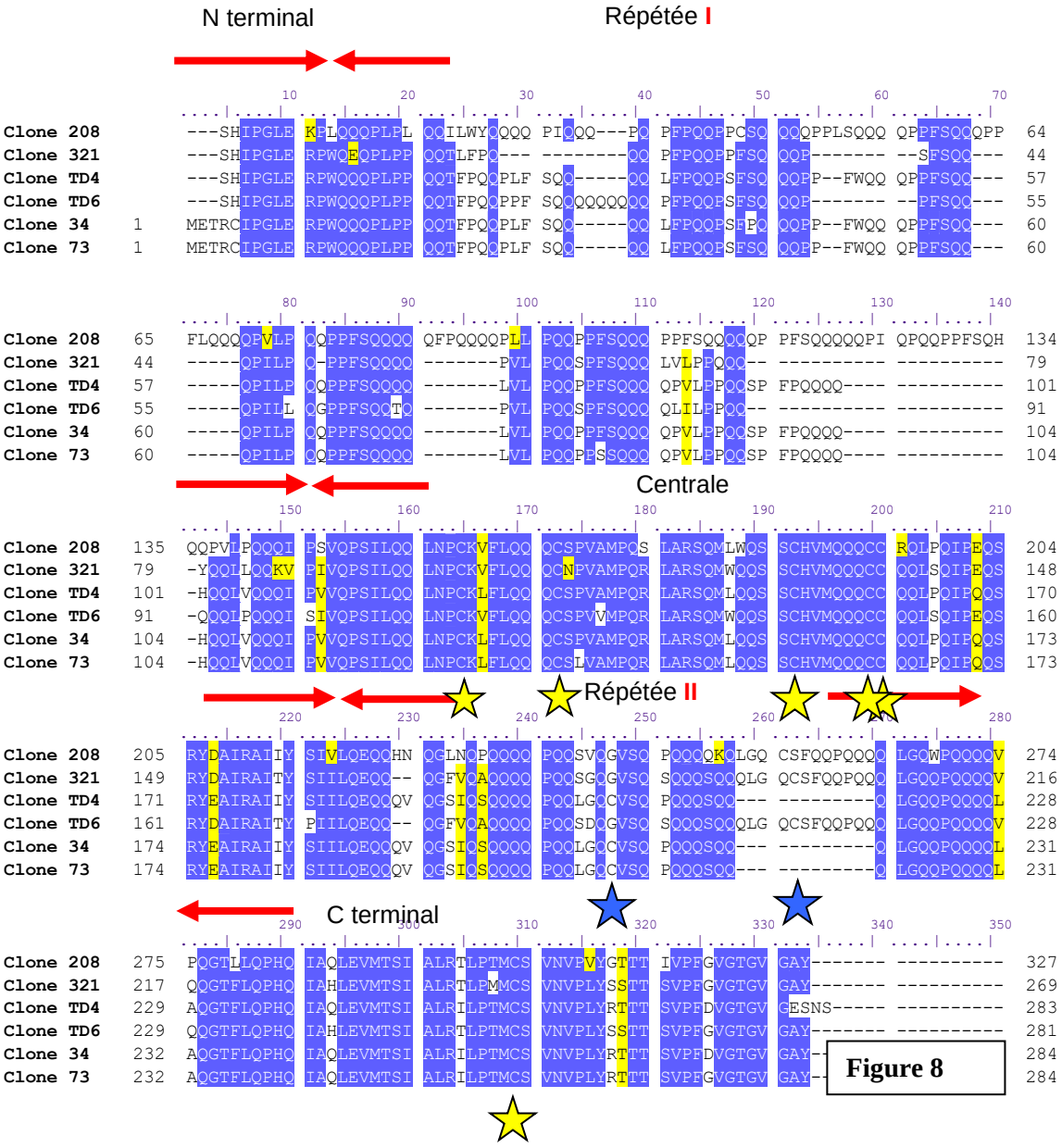


Figure 8

2.6 La transcription reverse

Le flux d'information génétique peut être dirigé de l'ARN vers l'ADN. Il existe une classe de virus, dont le matériel génétique est constitué d'ARN. C'est la cas par exemple du virus HIV. Il possède une enzyme spécifique, la reverse transcriptase qui est essentielle pour sa réplication. Cette enzyme multifonctionelle possède aussi une activité ADN polymérase ARN dépendante, ainsi qu'une activité de déplacement et de transfert de brin et une activité Rnase. Grâce à la reverse transcripase, le biologiste moléculaire pourra obtenir assez aisément un ADN, à partir d'ARN messagers.

III Biosynthèse et éléments de structure des protéines

3.1 Traduction

C'est le processus par lequel l'information contenue dans un ARNm va pouvoir donner naissance à une protéine. La traduction se déroule en 3 étapes : initiation, élongation et terminaison. L'efficacité de la traduction peut être affectée entre autres par la séquence de l'ARNm et par la fréquence des différents ARNt présents dans la cellule.

Le biais d'utilisation des codons

Le code génétique est dégénéré. Pour un acide aminé donné, il peut en effet exister plusieurs codons. Il en est de même pour les codons stop. Cependant on ne connaît qu'un seul codon pour le Trp (TGG), ainsi que pour la Met (ATG). Par contre la sérine peut être codée par 6 triplets différents. *In vivo,* les 6 ARNt transportant la Ser ne sont pas présents dans la cellule en quantité équimoléculaire. Les conséquences sur la quantité de protéine produite peuvent être importantes.

Si sur un ARNm, un seul codon est utilisé de manière majoritaire pour un acide aminé donné, et que l'ARNt correspondant à cet acide aminé est présent en quantités importantes, alors l'efficacité de la traduction sera grande. *A contrario*, si un codon codant pour un acide aminé est utilisé très fréquemment, et que l'[ARNt] correspondant est rare dans la cellule, alors la traduction sera bien moins efficace. On suppose que ce mécanisme permet de moduler la vitesse de la synthèse protéique, ce qui d'un côté joue sur la quantité totale de protéine produite, mais aussi pourrait laisser le temps à certaines régions / domaines des protéines de se replier correctement. L'indice de biais de codons (IBC) traduit donc une déviation par rapport à une utilisation au hasard des différents codons. L'IBC reflète un choix l'utilisation des codons codant pour un acide aminé. Il existe une relation entre l'IBC et la quantité de messager trouvé dans une cellule (Figure 9). La G3PDH, enzyme de ménage a un IBC de 0,99 elle représente 1,5 à 6 % des ARNm totaux. L'ARNm de l'iso 2 cytochrome C, dont l'IBC est proche de 0,15 est 500

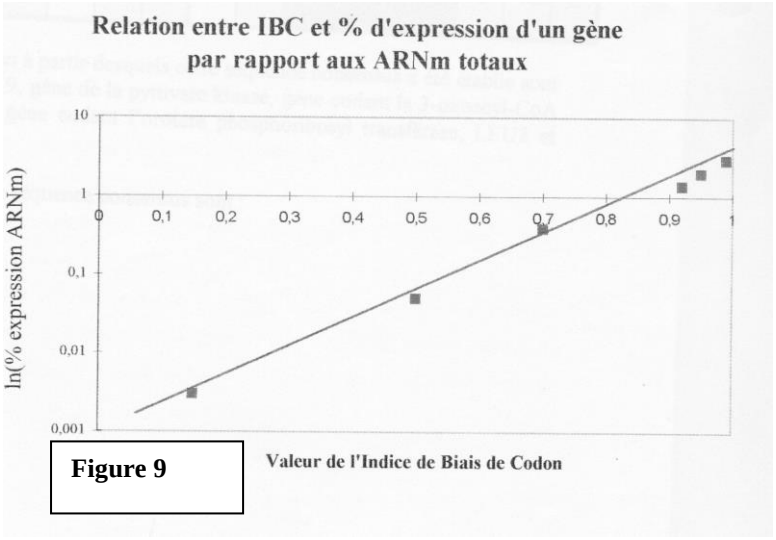


Figure 9

moins abondant. Cette relation a été décrite originalement par Benetzen ((*J Biol Chem* 1982, 257 : 3026-3031). Coghlan and Wolfe, (Yeast, 2000, 16: 1131-1145) ont suivi l'expression de tous les gènes du génome de *S. cerevisae,* à l'aide de puces à ADN. Ils ont montré qu'il existait une bonne corrélation (r^2 # 0,66, n=2525, $p<10^{-17}$) entre la concentration des ARNm trouvés dans la cellule et l'IBC.

3.2 Les acides aminés

Ce sont des molécules comprenant 1 groupement amine (NH2), 1 groupement carboxylique (COOH), et une chaîne latérale, de nature variée. Ces molécules possèdent un Cα qui est un centre chiral (sauf dans le cas de la glycine). Elles sont actives optiquement. On trouve habituellement 20 acides aminés dans les protéines. Ils appartiennent naturellement à la série L. On peut classer les acides aminés en trois groupes.

i Acides aminés à chaîne latérale non chargée à pH neutre

-Chaîne latérale aliphatique

Gly, Ala, Val, Leu et Ileu

Ser et Thr ont une chaîne latérale hydroxylée.

Cys et Met contiennent du soufre

La Pro a sa chaîne latérale cyclisée en noyau pyrole. Ce qui aura des implications importantes dans la structure des protéines.

-Chaîne latérale aromatique

Phe, Tyr, Trp

ii Acides aminés à chaîne latérale chargée + à pH neutre

Lys, Arg, His

iii Acides aminés à chaîne latérale chargée - à pH neutre

Asp, Glu (existent sous forme Asn, Gln)

On connaît beaucoup d'autres acides aminés, qui ne sont pas trouvés dans les protéines. L'incorporation d'acides aminés non naturels est une voie empruntée par certaines équipes de recherche pour modifier la fonction de protéines. De plus les acides aminé peuvent avoir des groupements additifs sur leurs chaînes latérales. On connaît plus de 150 modifications

différentes, parmi lesquelles la phosphorylation, la glycosilation, la cyclisation, les ponts di sulfure…..

3.3 La structure primaire :

C'est l'arrangement successif de résidus d'acides aminés dans une chaîne polypeptidique.

La liaison peptidique

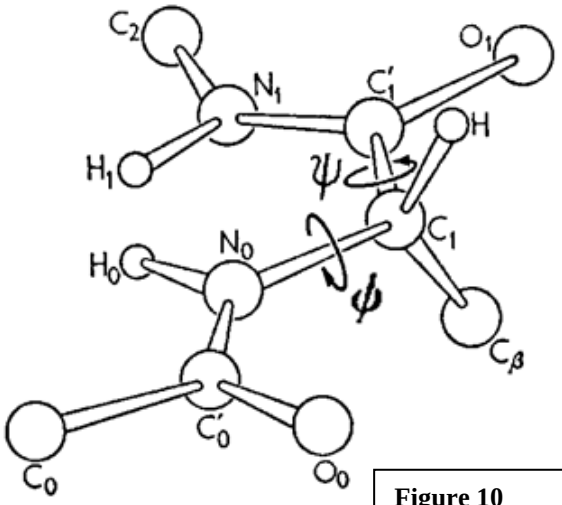


Figure 10

La liaison peptidique est une liaison amide, qui a un caractère partiel de double liaison. Les six atomes qui composent cette liaison sont compris dans un même plan. Cette liaison se comporte comme un dipôle, et une chaîne polypeptidique se présente donc comme un arrangement linéaire et flexible de plans rigides (unités peptidiques) pouvant pivoter autour des carbones α selon deux angles, ϕ et ψ (Figure 10). Ces angles sont appelés angles dièdres.

Suivant le nombre de résidus d'acides aminés qui les constituent, on distinguera :

Les oligopeptides de 2 à 10 résidus acides aminés. Ensuite, on parlera plutôt de peptides (10-50 résidus), puis de protéines (>50 résidus). Chez les organismes vivants on trouvera toutes les tailles de polypeptides, certaines protéines comportant plus de 6000 résidus d'acides aminés.

Les diagrammes de Ramachandran

La conformation locale d'un résidu peut être représentée par un point de coordonnées ϕ et ψ sur un diagramme en deux dimensions (Figure 11). Si l'on considère chaque atome comme une sphère dure ayant pour rayon le rayon de Van der Waals de l'atome considéré, certaines valeurs de ϕ et de ψ ne sont pas permises, car elles entraîneraient des collisions entre les atomes portés par le squelette

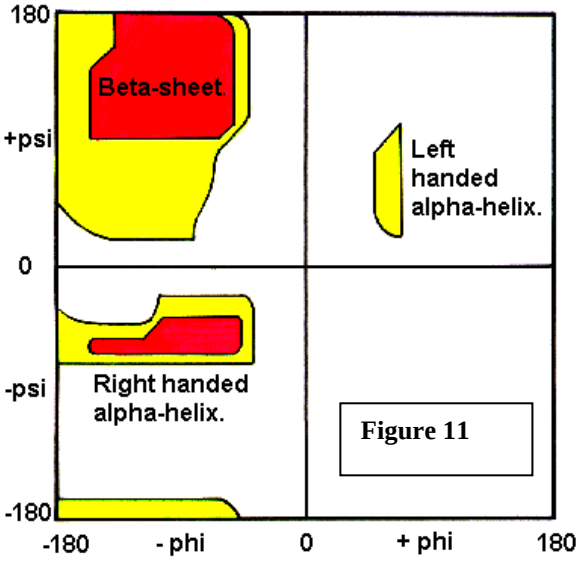


Figure 11

peptidique. L'analyse de structures tridimensionnelles de protéines à permis de constater que seules des zones correspondant à 8% du diagramme de Ramachandran sont « permises » quelles que soit la nature de la chaîne latérale. Ceci signifie que dans une protéine, les acides aminés ne peuvent pas adopter toutes les conformations que l'on pourrait attendre du fait de la présence de leurs chaînes latérales et aussi car ils sont fréquemment engagés dans des structures tridimensionnelles bien déterminées.

3.4 La structure secondaire

L'hélice alpha

Pour des valeurs de ϕ et ψ de –57° et de –47°, les acides aminés sont structurés en hélices alpha (Figure 12). Ces hélices sont stabilisées par des liaisons hydrogènes situées entre le CO d'une liaison peptidique et le NH d'une autre liaison peptidique située en direction du N terminal. Les chaînes latérales sont situées à l'extérieur de l'hélice. Une liaison peptidique comporte 3,6 d'acides aminés par tour. 1 tour de spire=0,54 nm. L'hélice α est stabilisée par des liaison H. L'hélice α dévie la lumière polarisée. Elle se comporte comme un dipôle électrique, ce qui lui permet par exemple de stabiliser un ligand chargé.

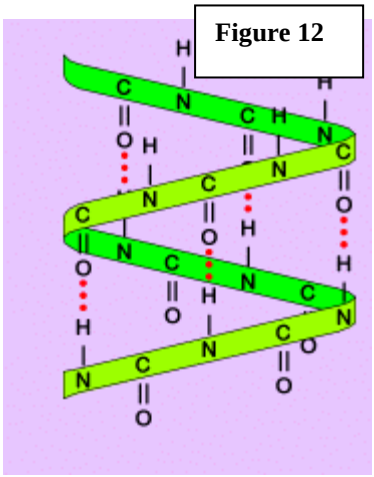


Figure 12

Le feuillet béta

Pour des valeurs de ϕ et ψ de –140° et de +135°, les acides aminés sont structurés en chaînes ß (Figure 13). Les chaînes ß peuvent former des assemblages dits en feuillets plissée (parallèles ou bien antiparallèles), stabilisés par des liaisons H inter caténaires (entre N peptidique et CO peptidique de brins différents). Les structures en feuillets ß ont une très forte tendance à s'agréger. Ceci est illustré par le cas de la protéine prion. Elle change de conformation. Elle passe d'une structure comportant 3 hélices α et 2

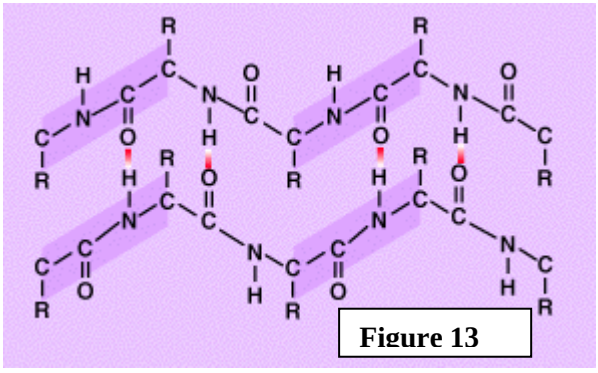


Figure 13

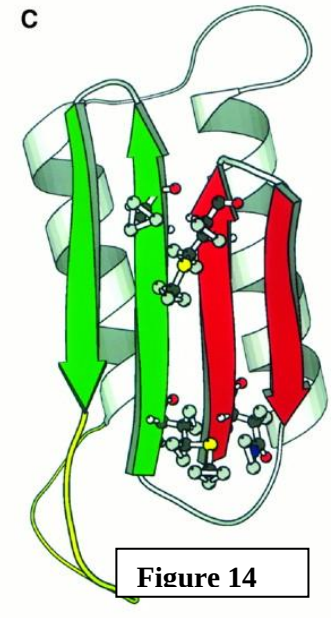


Figure 14

feuillets β à une structure 2 hélices α et 4 feuillets β. L'augmentation du contenu de la protéine en feuillet β rend cette protéine insoluble et résistante à la protéolyse (Figure 14).

3.5 La structure tertiaire

L'enroulement des protéines correspond à leur structure tertiaire, en particulier dans le cas des protéines globulaires, plus complexes que les protéines fibreuses. Attention, la différence entre structure secondaire et structure tertiaire est parfois arbitraire.

La conformation protéique est la résultante d'une multitude d'interactions faibles entre les différents atomes de la chaîne polypeptidique, et entre le polypeptide et son environnement. Chez les protéines globulaires, l'enroulement est compact. Il y a peu d'espace interne pour les molécules d'eau. (Attention, il peut cependant y en avoir quelques unes). Les résultats des recherches menées en RMN (Résonance Magnétique Nucléaire) et en radio cristallographie des protéines indiquent que les molécules de protéines sont entourées de une à deux couches d'eau. Dans le cas du lysozyme, la première couche d'eau (30 à 35 molécules) est très fortement liée, la seconde (95 à 105 molécules), est liée plus faiblement.

La structure tertiaire est essentiellement due à des interaction intra chaînes, entre des chaînes latérales d'acides aminés.

3.6 Dynamique des protéines

Notion de domaines

L'observation à l'aide de la radiocristallographie aux rayons X des protéines globulaires de masse supérieure à 20 kDa a montré que ces protéines étaient composées de structures repliées, liées entre elles par des liaison covalentes. Ces structures ressemblent à de petites protéines globulaires. Les domaines sont donc constitués de séquences protéiques qui forment des modules indépendants. Les domaines sont connectés par des régions relativement flexibles de la protéine. On peut considérer les domaines comme des unités fonctionnelles, caractérisées par exemple par leur capacité à fixer un cofacteur, ou bien un effecteur. Le domaine peut aussi être considéré comme une unité autonome de repliement. Cependant, deux domaines différents de la même protéine, isolés (ils n'appartiennent plus à la même chaîne polypeptidique) et repliés ne peuvent pas s'associer entre eux.

3.7 La structure quaternaire

Les protéines possédant plusieurs sous unités ont un niveau d'organisation supérieur : il s'agit de la structure quaternaire. Chaque chaîne peptidique d'une telle protéine est alors appelée

sous unité. On peut classer les sous unités en différents types, selon l'arrangement de leurs structures secondaires. On a aura ainsi :

des protéines α / α : c'est le cas de la myoglobine

des protéines α / β : c'est le cas de la flavodoxine

des protéines β / β : c'est le cas de la plastocyanine

des protéines α + β : c'est le cas du lysozyme

On dit qu'une protéine est oligomérique si elle contient deux sous unité ou plus.

La protéine kinase CK2 est une enzyme trouvée chez les organismes eucaryotes. Elle fait partie de la famille des protéines kinases, qui sont (sauf exception) des protéines régulatrices de l'activité des cellules. Ces protéines sont très conservées d'un point de vue structural. Leurs séquences comportent toutes 11 régions très conservées, et on connait plus de 2000 protéines kinases chez les eucaryotes. La protéine kinase CK2 phosphoryle *in vivo* et *in vitro* de nombreuses protéines en utilisant de l'ATP ou bien du GTP comme donneurs de phosphate. Cette enzyme possède deux types de sous unités : catalytiques (α) et régulatrices (β). La sous unité catalytique de la protéine kinase CK2 est divisée en deux domaines : l'un correspond à la partie C terminale, structurée essentiellement en hélice α (figure 15, partie inférieure), et l'autre correspond à la partie N terminale de la protéine qui est essentiellement structurée en feuillet β (Figure 15, partie supérieure, (D Chaillot, Thèse de l'INA PG, 1998). Ces deux domaines sont reliés par un coude, et montrent des mobilités différentes. Le site actif lui est situé à l'interface de ces deux domaines. Il ne représente qu'une toute petite partie de la protéine. Cette enzyme possède aussi une boucle riche en glycine (sans structure particulière, indiquée par une flèche), dont le rôle est de stabiliser les phosphates de l'ATP par l'intermédiaire de liaisons H.

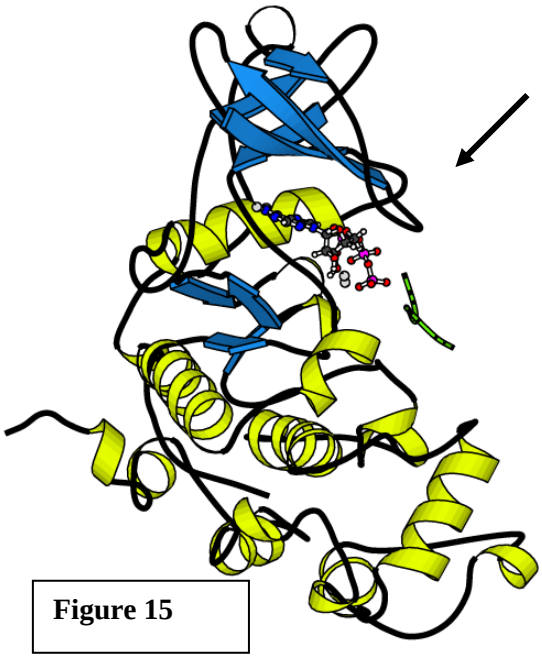


Figure 15

La sous unité catalytique de la protéine kinase CK2 (α) est active seule. Cependant, *in vivo*, on peut la trouver sous la forme de tétramères actifs ($\alpha 2/\beta 2$). La spécificité de l'enzyme tétramérique est différente de celle de l'enzyme monomérique.

Un atome de Zn, coordonné spécifiquement à 4 cystéines stabilise la conformation de la sous unité β, ce qui permet sa dimérisation (Figure 16, K Niefind 2001, *EMBO J*. 20, 5320-5331).

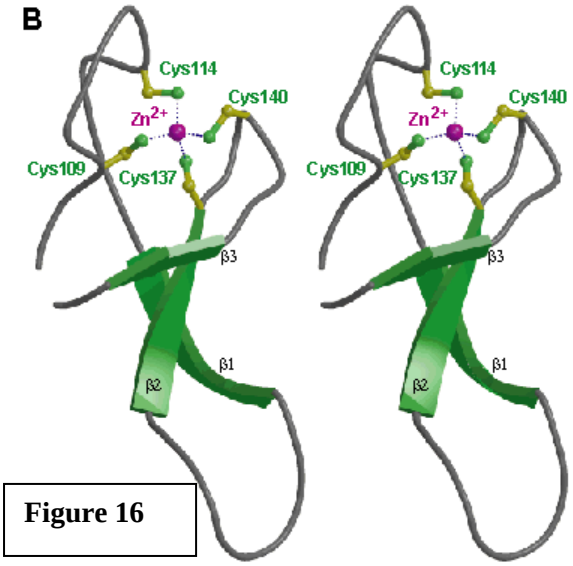


Figure 16

Alors, la protéine kinase CK2 peut s'organiser en tétramère (α2/β2). Celui ci est stabilisé par trois types de contacts.

Des contacts β / β: chaîne principale / chaîne principale

Des contacts α / β: chaîne principale / chaîne principale

Des contacts α / β: chaîne principale / extension de la sous unité β.

On a pu caractériser ces interfaces : aucun de ces contacts entre les sous unités n'est assuré par des liaisons covalentes. On peut calculer la surface de ces zones de contacts, et estimer leur qualité. Ainsi, on peut prédire si un oligomère sera « permanent » ou non.

La surface de contact β / β: # 1700 Å, forte composante hydrophobe, et liaisons H. La stabilité du dimère β / β sera très grande. Le dimère sera vraisemblablement permanent.

La surface de contact α / β # 800 Å est moindre que celle des contacts β / β. La stabilité de ces contacts sera certainement plus faible que celle des contacts β / β. La zone de contact α / β comprend une partie très conservée de la sous unité catalytique des protéines kinases. La liaison des sous unités régulatrices avec les sous unités catalytiques (α) est plus faible, d'où la possibilité de former des oligomères avec la sous unité catalytique d'autres protéines kinases. Le tétramère est vraisemblablement une entité non permanente.

La protéine kinase CK2 est donc une entité dynamique. La dynamique se trouve au sein des domaines (surtout dans cas de la sous unité α) qui bougent l'un par rapport à l'autre autour d'un coude, et dans le fait que les sous unité α peuvent se fixer ou bien quitter le dimère de sous unités β pour être vraisemblablement remplacées par d'autres sous unités catalytiques de protéines kinases.

3.7 Stabilité des protéines

La structure des protéines est stabilisée par (Figure 17)

- Les interactions hydrophobes entre chaînes latérales des acides aminés non polaires
- Les liaisons hydrogènes propres à la chaîne peptidique
- Les liaisons hydrogènes entre les chaînes latérales d'acides aminés
- Les liaisons ioniques entre les chaînes latérales d'acides aminés chargés (COO de Asp et NH3).
- Les ponts disulfures entre les résidus Cys.

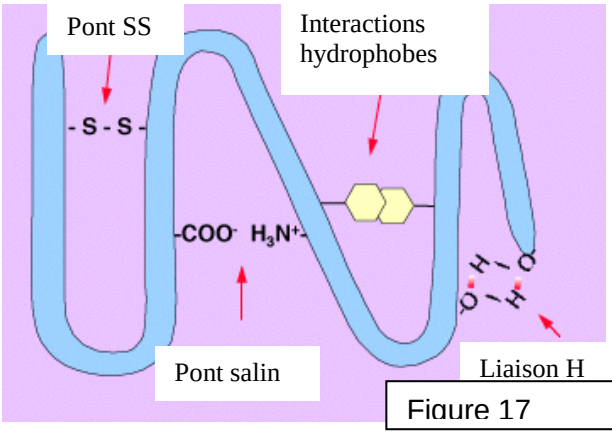


Figure 17

Les couches d'hydratation sont elles aussi extrêmement importantes

3.8 Dépliement repliement d'une protéine globulaire

On considère généralement que l'information nécessaire pour qu'une protéine se structure est contenue dans la séquence de ses acides aminés. Certains auteurs ont même écrit que le repliement d'une protéine était une deuxième lecture du code génétique. Le repliement des protéines est séquentiel et très rapide. Il se déroule en deux grandes étapes (Figure 18) :

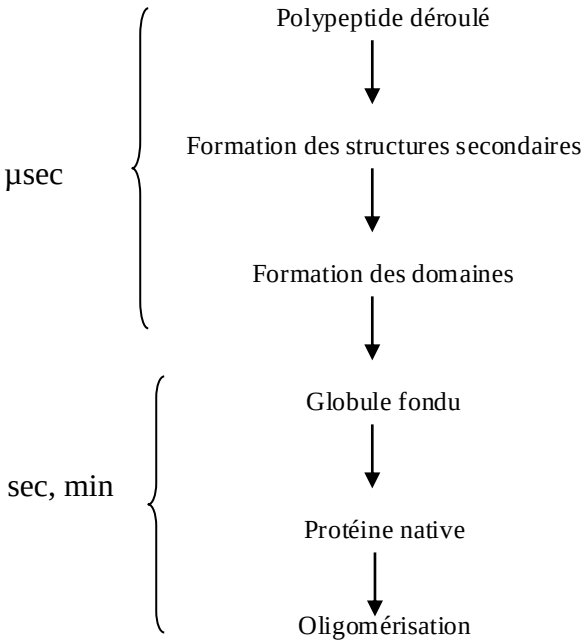


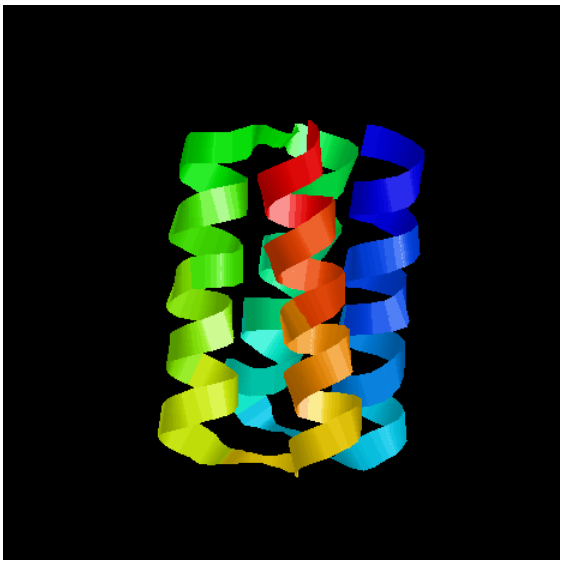
Figure 18

La chaîne polypeptidique dépliée adopte des structures secondaires. Ce processus est très rapide. On pense que le repliement de la protéine commencerait simultanément dans plusieurs secteurs de la chaîne polypeptidique, avec formation de courts fragments de structure secondaire. Le repliement serait donc orienté par les premières liaisons qui détermineraient les suivantes. Les structures secondaires se structurent en domaines, qui donnent naissance à un état particulier de la protéine: le globule fondu. Celui ci correspond a un état très souple de la protéine, qui permet l'arrangement final correct de la structure des domaines en une protéine native.

In vivo, le repliement est assisté par différentes protéines. Ce sont les chaperonnes, et les foldases. Les chaperonnes, qui appartiennent à la famille des protéines de choc thermique s'attachent reversiblement à des protéines mal repliées. Elles sont peu abondantes, et ont besoin d'ATP pour fonctionner. Leur faible abondance expliquerait pourquoi les protéines sur exprimées chez les bactéries, et aussi les levures sont fréquemment retrouvées sous forme de corps d'inclusions : il s'agit de protéines mal repliées, qui s'agrègent. Les prolines ont pour propriété de modifier l'orientation des structures secondaires dans lesquelles elles sont engagées. Pour la plupart des peptides, la structure entre le résidu n et n+1 est trans (ainsi les chaînes latérales ne se gênent pas). La forme trans est 1000 fois plus stable que la cis. Il est donc possible de trouver dans une protéine des prolines qui ne sont pas dans la bonne conformation. La peptidyl prolyl isomérase favorise l'isomérisation cis trans des prolines, ce qui permet d'obtenir une chaîne latérale correctement repliée. Les chaînes latérales cystéines sont fréquemment engagées dans des ponts disulfures. Ceux ci peuvent parfois être engagés entre des résidus qui normalement ne doivent pas être liés. Quant à la protéine disulfide isomérase, elle favorise les échanges de ponts disulfures, ce qui permet d'obtenir une protéine native possédant les ponts disulfure attendus. Les protéines peuvent être aussi modifiées post traductionellement. Ainsi elles peuvent être glycosylées, la méthionine N terminale peut être exisée, …. La structure des protéines est stabilisée par des interactions non covalentes, et aussi par des ponts dissulfure. Il est généralement possible de trouver des conditions permettant de déplier et de replier une protéine de manière à obtenir une protéine native. Il sera ainsi possible fréquemment à partir de corps d'inclusion, d'obtenir une protéine active. On utilisera pour cela des agents dénaturants (appelés aussi agents chaotropiques) tels que l'urée, le chlorure de guanidine, ou bien même un détergent, le sodium dodecyl sulfate, et l'on utilisera en même temps un agent réducteur. L'urée et le chlorure de guanidine établissent des liaisons hydrogène avec la liaison peptidique, et désorganisent la structure de l'eau au voisinage de la protéine. On obtient ainsi une protéine dépliée. Les agents réducteurs vont rompre les ponts disulfure. Il suffit généralement d'enlever les agents dénaturants (par dilution ou bien par dialyse) pour obtenir une protéine correctement repliée.

3.9 Des protéines repliées apparaissent fréquemment dans des bibliothèques de séquences aléatoires.

Qu'est ce qui fait qu'une séquence d'acides aminé se replie ? Parmi toutes les séquences d'acides aminés que l'on peut concevoir, les séquences repliées sont elles la règle, ou bien l'exception ? Ces questions sont apparemment très simples…
Sauer et ses collègues (PNAS 1994, 91, 2146-2150) ont construit une bibliothèque d'expression de gènes synthétiques codant pour des protéines de 80 à 100 résidus. Il s'agit d'une collection de plasmides dans lesquels on a inséré des séquences générées de manière « aléatoire », codant pour une succession de 3 résidus d'acides aminés . Les protéines exprimées seront ainsi composées majoritairement de répétitions aléatoires de glutamine (Q), de leucine (L) et d'arginine (R). Les gènes synthétiques sont construits de manière à contenir des codons Q (50%), L (40 %) et R (10%).
Glutamine et Leucine sont choisis comme acides aminés hydrophiles et hydrophobes. Quant à l'arginine, cet acide aminé chargé a été ajouté dans le but d'augmenter la solubilité des protéines.
On trouve dans ces bibliothèques d'expression 5 % des clones possédant la séquence codant pour la protéine QLR insérés dans un bon cadre de lecture correct qui expriment des protéines QLR. On peut les purifier. Ces protéines sont très peu solubles en absence d'agents chaotropiques. Ainsi QLR2 nécessitait la présence de plus de 4 M de Gdn HCL pour être soluble. Les trois protéines étudiées dans ce travail contiennent des parties repliées en hélice alpha et sont oligomériques.
Il est donc possible d'obtenir des protéines repliées, et même oligomérisées à partir de bibliothèques d'expression de gènes synthétiques. Ces protéines sont cependant différentes des protéines « naturelles : » Elles sont très résistantes à la dénaturation par le Gdn-HCl et par la chaleur. Elles restent structurées à 90°C en présence de 6,0 M de Gdn-HCL.
Leur très faible solubilité provient vraisemblablement de leur trop grande hydrophobie.
La méthode décrite ici est originale (au moins pour l'époque en 1994) car elle ouvrait la possibilité au moins théorique, de générer des protéines aux propriétés originales, codées par des séquences d'ADN « aléatoires », codant pour des protéines comportant peu de résidus d'acides aminés différents .
Hecht et ses collègues ont eux aussi essayé de créer un protéine structurée *de novo*.



La protéine Félix (Four helixes) (Figure 19, ci contre) a été crée *de novo* de manière à se replier en une gaine de 4 hélices α anti parallèles. Sa séquence comporte 79 résidus d'acides aminés, et ne présente aucune homologie avec les protéines connues à cette époque. Cette protéine est considérée comme "native" dans le sens qu'elle est non répétitive et qu'elle contient 19 des 20 acides aminés naturels. Felix a été exprimée à partir d'un gêne synthétique cloné dans *E. coli*. Cette protéine a été purifiée. Elle montre une bonne solubilité. Felix (i) est monomérique en solution, (ii) sa conformation est surtout en hélice alpha, (iii) elle comporte bien le pont dissulfure attendu entre la première et la quatrième hélice (iv) quant à son tryptophane unique, on le trouve bien dans un environnement apolaire. Il est donc ainsi possible de concevoir *de novo* une protéine à la conformation désirée. (Hecht MH, Richardson JS, Richardson DC, Ogden RC. Science 1990 Aug 24;249(4971):884-91).

Les deux idées développées par Sauer et par le groupe de Richardson consistaient à créer des protéines artificielles, i. e. dont les séquences d'acides aminés n'étaient pas retrouvées dans la Nature. Ils montrent qu'il est possible de concevoir des protéines dont on peut décider en grande partie de la structure. Nous verrons dans le dernier chapitre de ce cours que la bio diversité nous permet de trouver bien plus près de nous de nombreux exemples de protéines de propriétés définies. L'effort des chercheurs et des ingénieurs se portera sur l'amélioration des propriétés des protéines existantes plutôt que sur leur conception *de novo*.

IV Techniques de base de la biologie moléculaire : Clonage, expression homologue et expression hétérologue de protéines

4 .1 Principe

Il existe de nombreux systèmes d'expression de protéines, qui peuvent être mis en œuvre de manière plus moins complexe. Les opérations unitaires conduisant à l'expression d'une protéine suivront le même schéma général.

Tout d'abord il s'agira d'obtenir un maximum d'informations sur la séquence de la protéine que l'on cherche à cloner. La séquence codant pour cette protéine sera obtenue soit par criblage de banques d'ADN, soit par criblage de banques d'expression (dans ce dernier cas, on cherchera alors directement les bactéries qui possèdent l'activité recherchée (en suivant sur boîte le changement de couleur d'un substrat, l'apparition d'un halo….). Dans tous les cas, la séquence d'ADN sera insérée dans un vecteur d'expression. On vérifiera la structure du vecteur, et la séquence à exprimer. Ensuite, la phase d'expression sera lancée. Il faudra alors vérifier l'expression de la protéine, essayer de la purifier et ensuite la caractériser.

4.2 Production de protéines

On exprimera des protéines à l'aide de bactéries ou d'autres organismes

i Pour obtenir des quantités importantes d'une protéine homogène

ii Pour directement pour l'utiliser, ou bien pour étudier les propriétés de cette protéine

Les systèmes bactériens

Ce sont les systèmes les plus populaires. Le génome de différentes bactéries est connu (*E coli*)
Les bactéries sont faciles à utiliser, on possède de nombreux vecteurs d'expression. Les vitesses de croissance sont bonnes, et l'on peut obtenir des quantités de protéines relativement importantes. Cependant, les bactéries possèdent une machinerie permettant l'expression moins complète que celle des organismes eucaryotes. Les bactéries expriment très souvent les protéines étrangères sous forme de corps d'inclusion. Leur capacité de sécrétion est très limitée
Elle ne savent pas couper la Met N terminale, et leur capacité à faire des modifications post traductionnelles est très limitée (sauf en co exprimant l'enzyme responsable)
Les vecteurs d'expression de *E coli* parmi les plus utilisés sont ceux de la la série pET, qui utilisent un promoteur provenant du phage T7.

Les promoteurs T7 ne sont transcrits que par la polymérase de T7. Cette enzyme est 5 fois plus rapide que que la polymérase de *E. coli* pour l'élongation des ARNms. Aux fins d'expression dans *E. coli*, les gènes sont clonés en 3' du promoteur, et en 5' d'un site de coupure par la Rnase H, qui génère une stucture secondaire de l'ARNm, l'empêchant d'être dégradé.
Les vecteurs de la série pET (Figure 20) doivent être introduits dans des souches bactériennes qui possèdent un gène codant pour la polymérase de T7. Ce gène est sous le contrôle du promoteur *lac*, inductible par l'IPTG. Cependant, une quantité faible de T7 polymérase est présente, chez les cellules, en absence d'inducteur. Ceci peut gêner considérablement la croissance des cellules, surtout si l'on exprime des protéines toxiques (pour l'hôte). Pour éviter ceci, les cellules sont co transformées avec un plasmide qui porte le gène codant pour le lysozyme du phage T7. Le lysosyme du phage T7 est en effet un inhibiteur très puissant de la polymérase du phage T7. Ce lysosyme permet aussi de faciliter la lyse des bactéries après leur congélation en présence d'un détergent (le Triton X100). Les quantités de protéines produites par ce système sont de l'ordre de dix à quelques dizaines de mg / L de culture.

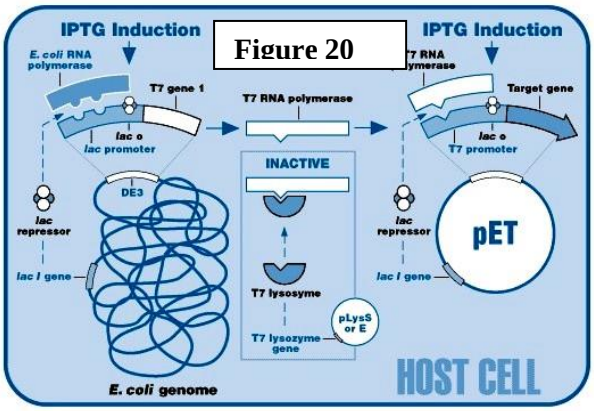
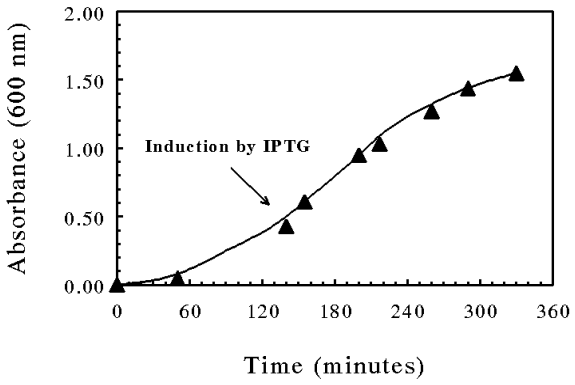


Figure 20

Nous avons utilisé la promoteur T7 pour exprimer la protéine kinase CK2 dans *E Coli*.



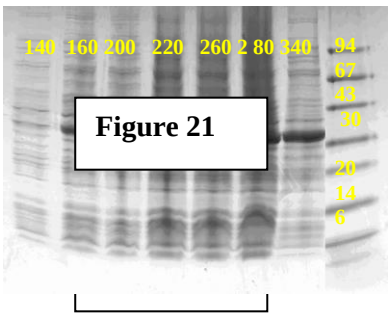
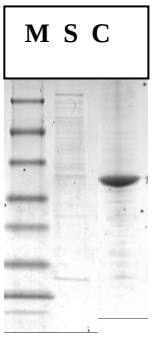


Figure 21



On fait croître les bactéries jusqu'à ce que la culture ait atteint une absorbance de 0,6 (figure 21 panneau de gauche). On ajoute alors au milieu de culture l'inducteur, puis on laisse les bactéries pousser encore 2 heures. On prélève régulièrement des échantillons de la culture sur lesquels on mesure les absorbances (panneau de gauche). On analyse par électrophorèse en milieu dénaturant ces mêmes échantillons prélevés régulièrement de 140 à 340 minutes, (panneau central). On voit une bande protéique apparaître aux alentours de 30 kDa (panneau du milieu) On récolte les bactéries, puis on les lyse. Le lysat est centrifugé, et on sépare alors le culot du surnageant . Les protéines sont alors

trouvées dans le culot sous forme de corps d'inclusion (panneau de droite, ligne C). En fait, une faible partie de la protéine existe sous forme soluble.

L'exemple suivant montre que l'on peut aussi diminuer la quantité des corps d'inclusion produits par la bactérie, en jouant simplement sur la température du milieu de culture.

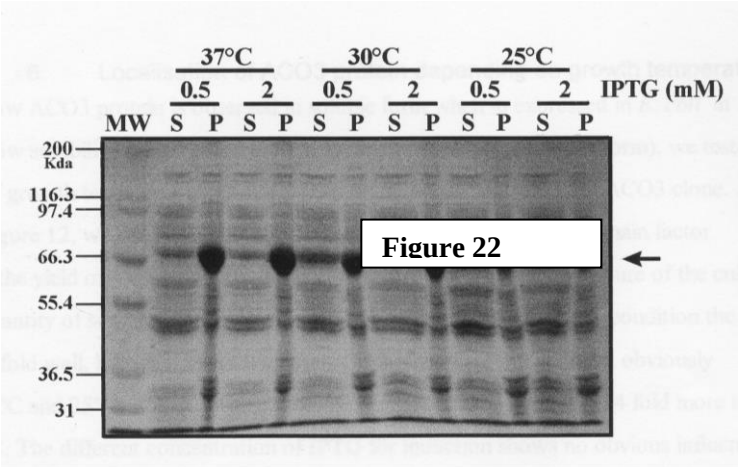


Figure 22

On voit ainsi dans la figure 22 que pour des températures de culture de 37°C, la protéine (dont la position est indiquée par une flèche) est majoritairement trouvée dans le culot (P), quelle que soit la concentration de l'inducteur introduite (IPTG) dans le milieu de culture. A 30°C, puis à 25°C, on voit une bande protéique dans les pistes de l'électrophorèse correspondant au surnageant : la protéine devient soluble (S)

Importance des biais de codons

Cette importance n'est pas toujours admise par les différents auteurs. En effet, quand on a des difficultés à exprimer des protéines, les explications potentielles de l'échec peuvent être très nombreuses (la protéine peut être toxique, être protéolysée par l'hôte, …).

Les ARNt reconnaissant les codons AGG AGA (Arg), AUA, CUA (Leu), CGA (Arg) et CCA (Pro) sont rares dans la bactérie. C'est ainsi que si l'on essaye d'exprimer dans *E. coli* certaines protéines (ici il s'agit d'une protéine allergène provenant de l'arachide), les quantités de protéines produites sont extrêmement faibles car la séquence codant pour cette protéine comporte de nombreux codons rares. On a cloné le gène codant pour cette protéine allergène dans deux systèmes d'expression différents. Tous utilisent un système T7. On a analysé par électrophorèse en milieu dénaturant (SDS PAGE) les protéines des bactéries transformées par un plasmide d'expression comprenant la séquence de la protéine allergène. La position à laquelle la protéine est attendue est indiquée par une flèche. Dans les panneaux du haut et du milieu de la figure 23, on ne voit pas la protéine attendue (ou bien elle est peu abondante). Si l'on utilise un autre système d'expression (Figure 21 panneau inférieur) comprenant des bactéries mutées de manière à exprimer certains des tARN rares alors

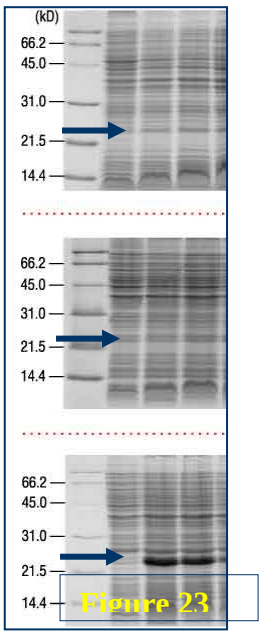


Figure 23

l'expression des antigènes est fortement améliorée. (Source Société Stratagène).

Les levures

La levure *S cerevisae* est elle aussi un système d'expression populaire. Son génome est connu. Elle est assez facile à utiliser. Les vecteurs utilisables pour exprimer des protéines sont nombreux. La synthèse des protéines peut être constitutive, induite. La levure croît relativement vite, elle peut donner une biomasse assez importante. Elle demande des conditions de culture peu onéreuses. Elle est capable de réaliser des modifications post traductionelles qui ne sont pas possibles chez les bactéries. Finalement, elle est plus efficace que les bactéries pour sécréter les protéines, mais *S cerevisae* ne sait pas bien enlever la Met N-terminale, et à tendance à hyperglycosyler les protéines.

Il existe des levures qui constituent des alternatives à l'utilisation de *S cerevisae*. La levure *P pastoris* , commercialisée sous forme de Kits montre de bonnes capacités de sécrétion. On recense plus de 200 protéines différentes exprimées avec succès par cette levure, les rendements allant de quelques µg / L à plusieurs g / L. (JL Cereghino and JM Cregg, *FEMS Microbiol Let.t,* 2000, 24 p 45-66).

La levure *Y. lipolytica*, étudiée à Grignon par le laboratoire de C Gaillardin est un excellent outil de sécrétion de protéines elle aussi.

Le génome de cette levure non pathogène (on la trouve naturellement sur certains fromages) est presque entièrement séquencé. Elle est assez facile à utiliser, plusieurs vecteurs d'expression sont d'ores et déjà disponibles, permettant une expression induite, constitutive, ou pouvant être réprimée. Il est possible d'intégrer ces vecteurs soit sous forme unique, soit sous forme multicopie. Elle croit relativement vite, et conduit à l'accumulation de biomasses importantes. Elle sécrète naturellement une lipase et une protéase alcaline. Cette levure permet les modifications post traductionelles, possède une très bonne capacité de sécrétion. Des souches ne sécrétant pas de protéases existent.

Certaines protéines sont exprimées à des niveaux proches du gramme par litre .

Les plantes

Les plantes sont des systèmes d'expression très intéressants car, elles sont généralement faciles et peu onéreuses à cultiver. Ce sont des eucaryotes qui peuvent faire de nombreuses modifications post traductionnelles. Plus encore, on sait cibler les protéines vers un compartiment donné (racines, feuilles, graines). Les seuls défauts que l'on peut leur trouver sont celui de l'acceptabilité des plantes transgéniques, et d'éventuels problèmes d'allergénicité, et aussi du transfert de gènes étrangers à des plantes. Les graines qui peuvent résister à la dessiccation stockent leurs lipides de réserve sous forme de corps lipidiques. Ce sont des globules, composés d'un cœur de lipides, entouré de phospholipides et protéines : Parmi celles ci, les oléosines sont intéressantes car elles sont accrochées fortement au corps lipidique, et elle sont abondantes. Pour purifier les corps lipidiques, il suffit de broyer les graines, et de centrifuger le broyat . Les corps lipidiques sont récupérés par simple flottation.

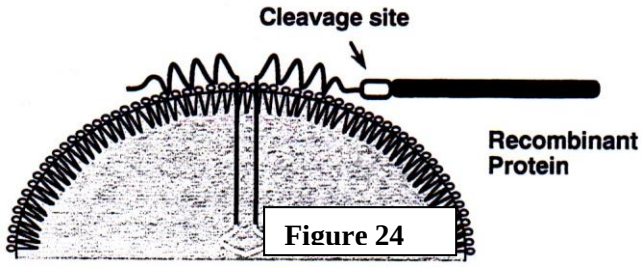


Figure 24

Les chercheurs de l'université de Calgary ont eu l'idée de cloner aux fins d'expression des protéines sous forme de chimères avec les oléosines, et sous le contrôle du promoteur des oléosines, qui est un promoteur fort. L'introduction entre la séquence de la protéine à exprimer et la séquence des oléosines d'un site de coupure par une protéase (Figure 24, Source Société SemBioSys) permet de récupérer facilement la protéine étrangère.

Ce système présente plusieurs avantages : la plante possède des mécanismes d'expression de protéines évoluées. Le ciblage de l'expression vers la graine permet de concentrer la protéine, et de la stocker sous une forme peu hydratée. On est seulement limité par la superficie du champ !! Bien sûr, la conduite de la culture d'un champ est plus simple que celle d'un fermenteur !! (Quoi que…. ?)

L'hirudine est un polypeptide comprenant 65 acides aminés, que l'on trouve naturellement dans la salive de la sangsue. Cet anticoagulant présente plusieurs avantages par rapport aux anticoagulants de la famille de l'héparine, parmi lesquels une bien meilleure capacité à inhiber la thrombine (dernière enzyme de la cascade de la coagulation du sang). Comme il est illusoire de purifier l'hirudine à partir de la salive de sangsue, on essayé de cloner et d'exprimer cette enzyme dans différents organismes. Chez *E coli*, la protéine obtenue est inactive, car elle possède toujours sa Met N terminale. La protéine a été clonée sous forme d'une protéine de fusion oléosine - hirudine. Elle est correctement exprimée *in planta*. La chimère hirudine oléosine

est correctement insérée dans l'oléosome. On la purifie « simplement » à partir d'oléosomes obtenus par broyage des graines. Les oléosomes flottent à la surface d'une solution aqueuse. On les soumet alors à une protéolyse afin de libérer l'hirudine. Cette protéine devrait être commercialisée sous peu au Canada.

Voici quelques exemples de protéines exprimées par des plantes

On peut les classer en 3 grandes classes : les vaccins, les anticorps, les molécules d'intérêt pharmaceutique ou bien biotechnologique. En voici quelques exemples (en 2000, on recensait au moins 40 protéines différentes produites dans des plantes).

	Plante	**Protéine**	**Vecteur**	**Firme**
Vaccin	Tabac	HbsAG (antigène de surface de l'hépatite B)	Agrobacterium	
	Pomme de terre	CtoxA et CtoxB (vibrion cholérique)	Virus mosaique de la pome de terre	
Anticorps	Tabac	Anti corps anti créatine kinase	Agrobacterium	
	Céréales	Anti corps ant anti carcinome embryonnaire	Bombardement de particules	
Molécules d'intérêt biotechnologique	Tabac	Hémoglobine	?	Meristem Therapeutic
	Tabac	Lactoferrine humaine	?	Meristem Therapeutic
	Tabac	Lipase acide de chien	?	Meristem Therapeutic ?

Source Nature Biotechnology November 2000, page 1151-1155

Les cellules de mammifères

Elles semblent être les candidates idéales pour l'expression de protéines car elle sont capables de faire les modifications post traductionelles, on peut les cultiver à grande échelle. Cependant les rendements en protéines sont faibles, le coût des matériels et des produits est élevé. De

plus, les vecteurs utilisés pour transformer ces cellules sont potentiellement pathogènes : Ce sont soit des virus, soit des oncogènes

L'araignée possède 7 glandes qui sécrètent des fibres protéiques différentes (Figure 25, d'après Vollrath *et al, Nature* 2001 ; 410, p541-548)). Certaines fibres servent à envelopper l'œuf de l'araignée, d'autres servent à capturer la proie. D'autres servent à accrocher la toile, à servir de fil de traction pour l'animal. Les protéines de la toile d'araignée bien qu'elles soient filées a pression et températures ambiantes, et en utilisant de l'eau comme solvant ont des propriétés mécaniques qui sont peu différentes de celles de l'acier….pour un poids # 8 fois plus faible. Cette soie est le fruit de 400 millions d'années d'évolution, elle permet d'arrêter une proie volante, et de l'immobiliser. En général, ces protéines sont constituées de motifs répétés. Elle contiennent peu de lysine et d'histidine, et aucune cystéine. Elle sont assez riches en Ser, Gly, Ala. Les fils de traction comportent de nombreuses Alanines trouvées dans des blocs répétés (ASAAAAAA) qui donnent aux fibres leur propriétés mécaniques, des blocs riches en glycine [(GGYGPG), ou bien (GPGQQ)n] qui sont responsables de l'élasticité aux filaments. L'araignée arrive à filer cette fibre par un judicieux contrôle du repliement et de la cristallisation des composants protéiques. Ces protéines ont donc suscité l'intérêt des scientifiques depuis de nombreuses

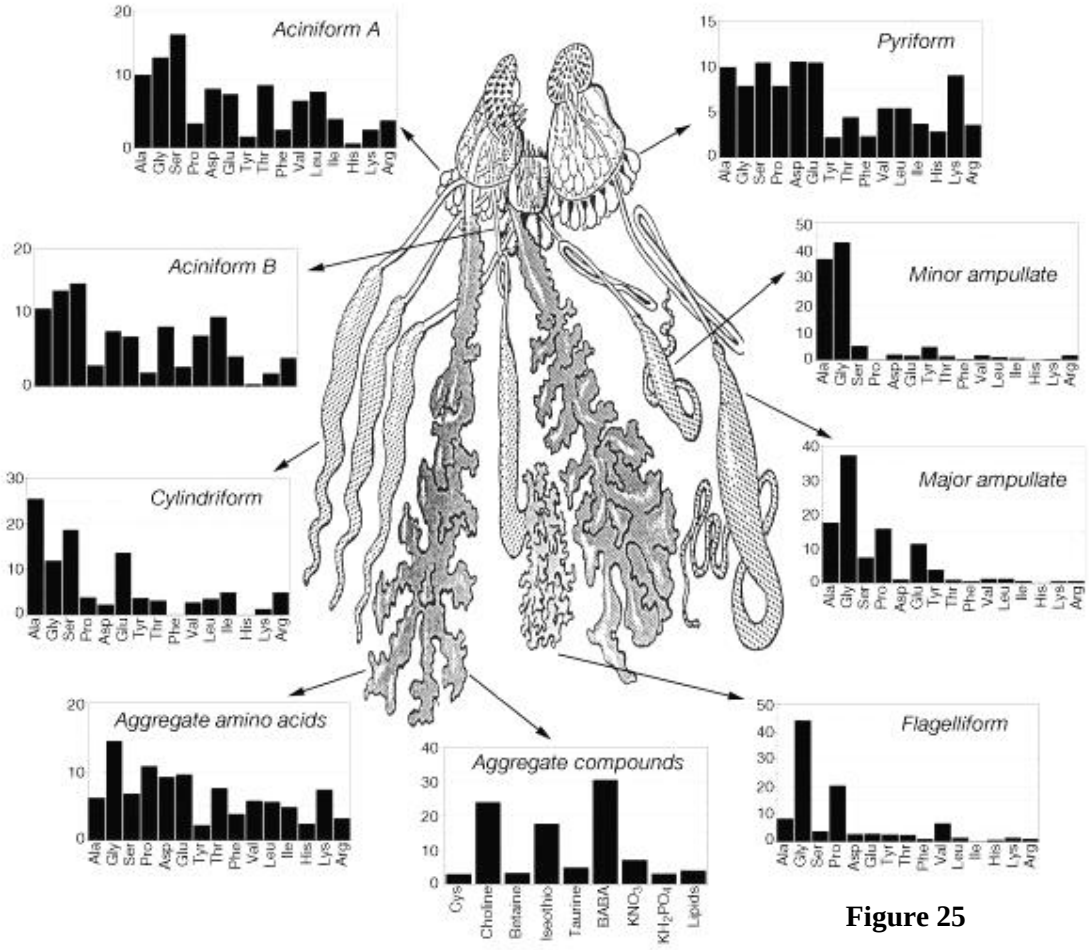


Figure 25

années. On a ainsi produit avec un succès limité ces protéines chez *E. coli*, chez la levure, et même chez des plantes. Du fait de la structure très répétée des séquences, les ARNm on des structures secondaires, ce qui gène leur traduction correcte. Le caractère très répété de la séquence de l'ARNm implique que les pools d'ARNt correspondant soient importants pour que la traduction soit efficace, ce qui est le cas chez l'araignée…et pas chez les bactéries ou bien les levures. Quand aux groupes qui ont essayé de filer ces fibres, ils ont du solubiliser les protéines en utilisant des systèmes de solvants assez agressifs (hexafluoro isopropanol), et ont obtenu des fibres aux propriétés bien moins bonnes que celles des fibres natives.

L'équipe de Lazaris (Science Janvier 2002, 295, 472-476) a utilisé des cellules de mammifère, afin de tester leur potentialité pour obtenir des fibres filables de bonne qualité. Ces cellules ont été utilisées pour sécréter des protéines de la toile d'araignée (fibroines). Ils ont obtenu une expression correcte de différentes protéines, parmi lesquelles une protéine dont la masse était proche de 60 kDa (ADF3). Ces protéines ont été précipitées à l'aide de sulfate d'ammonium, puis ont été re suspendues dans du tampon phosphate salin (PBS). ADF3 était soluble dans le tampon PBS, ce qui n'était pas le cas quand cette protéine a été exprimée dans *E. coli* ou bien par des levures. Ceci vient certainement du fait que la protéines exprimée dans *E. coli* ou bien dans la levure étaient tronquée (il lui manquait un fragment COOH terminal hydrophile). Ces protéines ont ensuite été filées au moyen d'un appareil « bricolé » pour la circonstance.

Si on file les protéines dans l'eau, on peut obtenir des fibres de 8 à 40 µm de diamètre, qui auront des propriétés mécaniques plus ou moins bonnes. On étirera ensuite les fibres de plus grand diamètre jusqu'à ce que leur diamètre soit de l'ordre de 20 µM. On peut ainsi les étirer 5 fois. On les laisse ensuite sécher en les maintenant attachées par leurs extrémités. Les fibres ainsi obtenues ont des propriétés mécaniques tout à fait comparables à celles fabriquées par les araignées.

Il est donc possible de fabriquer des fibres filables, possédant de bonnes propriétés mécaniques en utilisant des protéines de soie d'araignée sécrétées par des cellules de mammifères.

Ces protéines sont en apparence très « simples » car elles sont constituées de peu d'acides aminés organisées selon des séquences répétées, cependant, on ne sait encore pas les filer aussi bien que l'araignée.

V La diversité naturelle, la diversité générée

Les molécules qui composent les êtres vivants sont le résultat de millions d'années de mutations et de sélection. Les adaptations que l'on observe à l'échelle moléculaire ne sont pas non moins remarquables que celle que l'on peut observer à l'échelle de l'organisme.
De plus, c'est seulement maintenant que nous commençons à prendre conscience que la biodiversité qui nous entoure semble être un réservoir de molécules quasi infini. Ainsi, l'homme a récemment exploré des biotopes aussi différents que les fumeurs noirs que l'on trouve au fond des océans, des sources d'eau chaude sulfureuses situées au fond des geysers, dans des gisements pétroliers. On a trouvé dans ces différents biotopes des collections de micro organismes adaptés aux conditions extrêmes de pH, de température, et de salinité. Ces collections de micro organismes, que l'on appelle des extrêmophiles sont une source d'enzymes capables de fonctionner dans des conditions extrêmes de température, de pH et de salinité. Plus près de nous, des milieux familiers tels que les caves d'affinage de fromage au lait cru ou les compost livrent des flores microbiennes riches de plusieurs centaines de souches différentes, dont la diversité génétique est accrue par la présence de nombreux éléments transposables.
Depuis une dizaine d'années, les biologistes ont entrepris de diriger l'évolution des molécules. Il s'agit de mimer les mécanismes moléculaires de l'évolution afin d'améliorer la fonction d'enzymes. Pour cela, on utilisera principalement des méthodes favorisant l'apparition des erreurs, et celles permettant la recombinaison de molécules d'ADN entre elles. Il s'agit de générer la diversité. Tout l'art consistera ensuite à isoler les variants d'intérêt parmi les très nombreux variants qui auront été générés.

5.1 Comment faire apparaître des erreurs ?

Il est possible de muter une séquence d'ADN en utilisant des techniques physiques (rayonnements ultraviolets par exemple), ou chimiques (la liste des agents mutagènes est très longue). Les mutations seront réparties de manière « aléatoire » sur l'ADN. Il est aussi possible de muter une protéine d'une manière raisonnée, en introduisant les changements dans des zones précises, dont on connaît , ou dont on veut mieux connaître la fonction. Il s'agit alors de mutagenèse dirigée. Ces techniques sont maintenant assez classiques, et elles sont bien décrites dans de nombreux ouvrages. Aussi je préfère montrer comment il est possible de générer de la diversité afin de la combiner à la bio diversité pour d'obtenir des protéines aux propriétés recherchées.

La méthode de choix pour obtenir un ADN est bien sûr la PCR. Il existe différentes polymérases thermostables sur le marché. Elles font toutes des erreurs, de 1 x 10-4 à 2 10^{-5} erreurs / paire de bases pour la Taq de *Thermus aquaticus*, à 1.6 x 10-6 erreurs / paire de bases pour la Pfu polymérase (*Pyrococcus furiosus*). Si l'on veut obtenir de nombreux variants d'une séquence codant pour une protéine, on évitera d'utiliser une polymérase ayant une activité correctrice. Au contraire, on va jouer sur les conditions dans lesquelles la réaction de polymérisation se déroule pour augmenter et essayer de contrôler le taux d'erreur. Idéalement, il faudrait obtenir une population d'ADN portants chacun une mutation unique. Il est donc possible de prendre une polymérase qui fait des erreurs, et de la faire travailler dans des conditions qui vont amplifier ces erreurs. On jouera ainsi sur la concentration des dNTP, du chlorure de magnésium, sur le pH de la réaction. Ainsi, si l'on passe d'une concentration en Mg de 1 mM à une concentration de 20 mM dans tube de réaction de PCR, le taux d'erreur va passer de 2 10^{-5} à 210^{-4}.
On pourra aussi jouer sur les concentrations relatives des 4 dNTPs présents dans le milieu de réaction.
On peut donc générer aisément par PCR une population constituée d'ADN portant chacun une mutation (ou du moins on l'espère).

5.2 Comment accumuler les mutations favorables au sein d'une même molécule d'ADN entre elles ?

Si l'on examine deux enzymes homologues appartenant à deux organismes, l'un étant thermophile, l'autre mésophile , nous remarquerons que les séquences de ces enzymes peuvent différer sur plusieurs positions. Il est donc très difficile de changer les propriétés d'une enzyme à l'aide d'une seule mutation. Il va falloir trouver un moyen « prendre le meilleur » de toutes les séquences de la population d'ADNs portant chacun une mutation que nous avons générée. La technique la plus simple (Figure 26) (D Pompon et al Biofutur 2001 ; 214 ; 40-46) sera de couper les ADN ainsi obtenus par des enzymes de restriction, de purifier les fragments, et de les

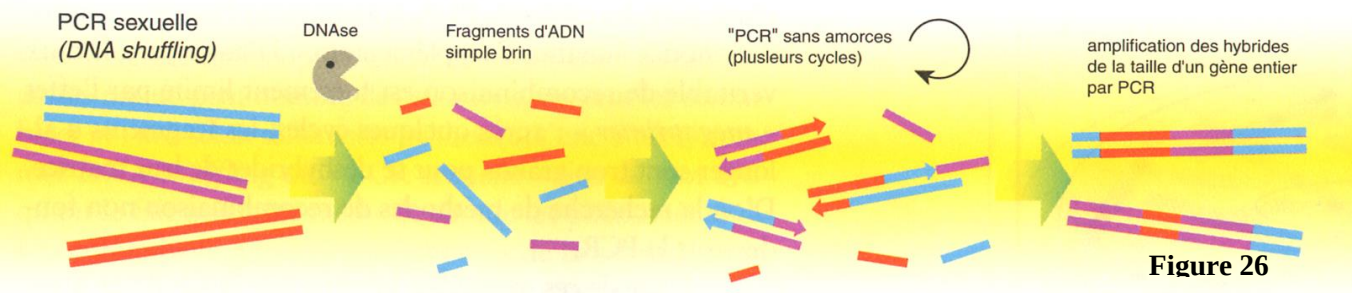


Figure 26

amplifier par PCR. Nous obtiendrons ainsi des fragments d'ADN chimériques que nous pourrons

insérer dans les vecteurs d'expression. Ces vecteurs d'expression serviront à transformer des bactéries au sein desquelles nous rechercherons les enzymes voulues au moyen de cribles. Il est important de bien comprendre que parmi toutes les molécules d'ADN que nous avons ainsi fabriquées, seules une petite population présentera un intérêt. Ces techniques sont très puissantes, mais il faudra à tout prix posséder des robots de criblages pour pouvoir trouver le bon mutant dans des populations qui peuvent très facilement dépasser le milliard d'individus (rappelez vous du nombre de molécules d'ADN que l'on peut fabriquer à partir d'un ADN en utilisant la PCR…). Il faut aussi que le crible soit « facile » c'est à dire qu'il faut un substrat stable, pas cher, qui change de couleur sous l'action d'une enzyme…ce qui est assez rare !

5.2 Cas de l'alpha amylase

Le fractionnement humide du maïs est un processus industriel multi étapes, qui fait intervenir des enzymes

Une des premières fraction d'amidon obtenue a un pH de 4,5. cet amidon est ensuite liquéfié à partir de l'amidon semi purifié en oligomères de glucose par l'alpha amylase de *B. licheniniformis*., cette étape se faisant idéalement à pH 4,5 et à 105 °C. Comme l'enzyme est instable dans ces conditions, il faut augmenter le pH à 5,7 6, et ajouter du calcium. La seconde étape est une saccharification du produit liquéfié qui utilise une gluco amylase d'*A species*, dont le pH optimum est proche de 4,2 - 4,5. Il faut donc réajuster le pH. La dernière étape consiste en la conversion du glucose saccharifié en un sirop de fructose, en utilisant la glucose isomérase. Avant cette étape, il faut enlever le Ca, et les sels qui proviennent de l'ajustement des pH.
Cette étape est coûteuse, et elle serait évitée si on utilisait une alpha amylase qui ait un pH optimum de 4.5, une température optimale 105 °C, et soit stable en absence de Ca.
C'est cette alpha amylase « idéale » que les chercheurs de la société Diversa ont recherchée. Ils ont tout d'abord fait des prélèvements de matériel biologique dans un certain nombre de biotopes (qu'ils ne révèlent pas). A partir de ces prélèvements, ils ont purifié des ARN, à partir desquels ils ont fabriqué des ADNc. Ces ADNc ont été clonés dans des plasmides permettant l'expression de protéines dans des bactéries. Les bactéries transformées à l'aide de ces plasmides ont été étalées sur des milieux contenant un antibiotique (afin de ne garder que les bactéries possédant le plasmide qui confère une résistance à un antibiotique), et de l'amidon (pouvant disparaître en donnant un halo si de l'alpha amylase est exprimée par les bactéries). Ils

ont gardé alors les bactéries qui poussaient sur boite et qui exprimaient une alpha amylase. Seuls 15 clones parmi les 50000 exprimaient une alpha amylase. Ils ont alors cultivé les bactéries exprimant ces alpha amylases en milieu liquide, induit la synthèse des différentes protéines. Les enzymes ont alors été purifiées, et on a étudié leur activité enzymatique dans différentes conditions (activité, pH, température, sensibilité à la présence de calcium…). Chaque enzyme a pu être caractérisée (pH optimum, température optimale). L'ADN de ces enzymes a été séquencé . Aucune enzyme ne satisfaisait aux trois critères requis (pH optimum de 4,5, température optimale de 105 °C, enzyme stable en absence de Ca) simultanément.

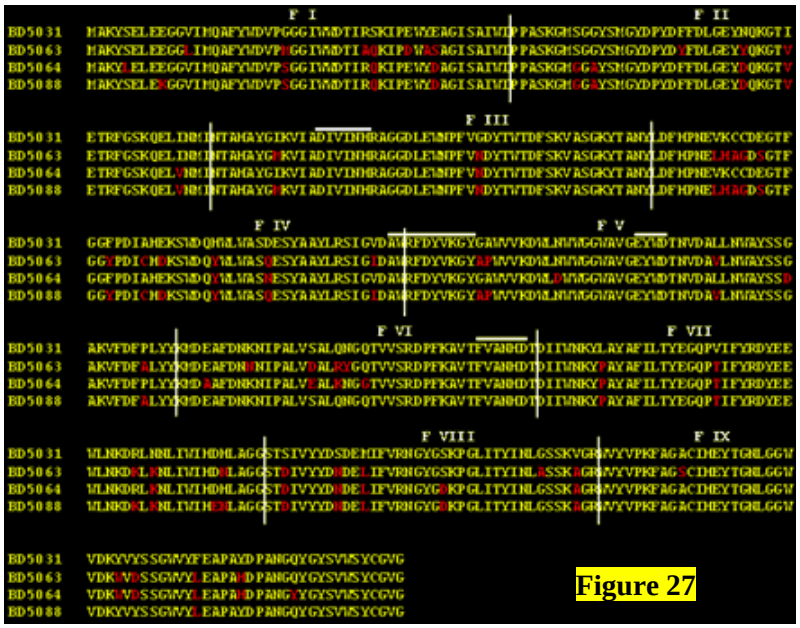


Figure 27

Les chercheurs de Diversa (Richardson *et al.* J Biol Chem 2002, 277, 26501-26507) ont décidé de combiner les séquences codant pour 3 alpha amylases afin d'en générer une qui possède simultanément les trois propriétés recherchées (pH optimum 4,5, température optimale 105 °C, enzyme stable en absence de Ca). Il s'agissait de « prendre le meilleur » chez ces différentes enzymes. L'ADN des ces bactéries a été coupé par des enzymes de restriction, de manière à exciser la séquence codant pour l'enzyme recherchée. Ces séquences codantes ont été coupées à l'aide d'enzymes de restriction au sein de zones d'homologie en 9 fragments (figure 27), ce qui donne 3^9 possibilités de recombinaisons. Elles ont été assemblées en séquences codantes chimériques. Ces séquences codantes ont été clonées dans un plasmide permettant l'expression de protéines. Ces plasmides ont servi à transformer des bactéries. On a obtenu 20000 clones. L'activité alpha amylase a été trouvée dans # 40 % des clones examinés (soit plusieurs centaines). On a trouvé 34 clones qui avaient une meilleure activité à pH 4.5 que l'alpha amylase de *B. licheninformis* (figure 28). 22 de ces clones avaient

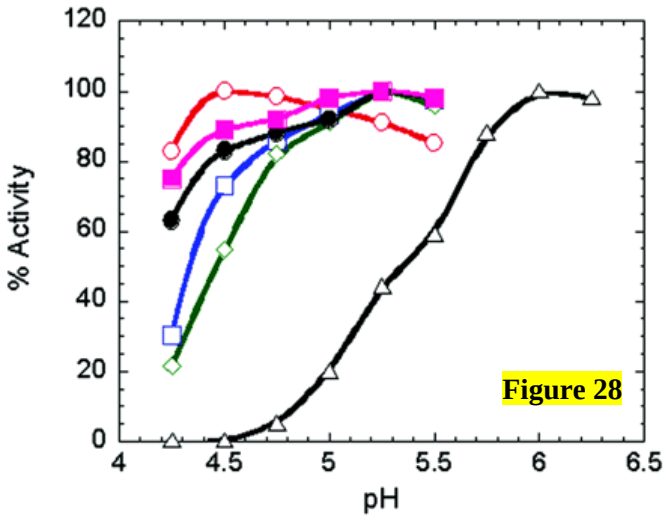


Figure 28

une meilleure thermostabilité à 90°C que les enzymes parentes en absence de Ca ajouté au milieu de réaction (la demie vie de certaines de ces enzymes a été augmentée d'un facteur 40 par rapport à l'alpha amylase de *B. licheninformis*) . 20 de ces clones avaient une meilleure thermostabilité à 100°C après ajout de Ca au milieu de réaction que les enzymes parentes. Certaines de ces enzymes ont été testées dans des conditions industrielles à l'échelle pilote. Les produits de la réaction d'hydrolyse de l'amidon par les alpha amylase recombinées sont similaires à ceux obtenus par l'action de l'alpha amylase de *B. licheninformis* sur de l'amidon.

Il est donc possible d'améliorer significativement des enzymes en utilisant les techniques modernes de la biologie moléculaire. On utilisera donc soit la bio diversité si on en dispose, soit une diversité générée afin d'obtenir un maximum de variants d'une enzyme. On teste les propriétés d'un grand nombre de ces variants , et on recombine les variants entre eux de manière à « tirer le meilleur » de cette population. Attention, ces techniques sont très esthétiques, mais elle nécessitent des systèmes de crible performants (robots, vision artificielle, etc…). Si il est assez facile de détecter une activité alpha amylase , lipase ou bien protéase (il existe des substrats synthétiques de cette enzyme qui changent de couleur après hydrolyse), la mise en évidence d'activités enzymatiques n'est pas toujours aussi simple.

Par ces techniques, il est aussi possible d'envisager d'améliorer les propriétés de protéines qui ne sont pas des enzymes : des anticorps, des récepteurs…

Pour en savoir plus :

Sur l'ingénierie des protéines en général

L'ingénierie des protéines et ses applications Editions Lavoisier Tec et Doc Ouvrage collectif coordonné par H Heslot 1996 ISBN 2-85206-991-1

Chapitre I à III

Biochemistry, Rawn

Chapitre IV

Cours de M Rochet, MC AgroParisTech

PCR a practical Approach The Practical Approach Serie IRL Press ISBN 0-19-963226-X

Molecular Cloning, a laboratory manual. Sambroook, Fritsch and Maniatis Cold Spring Harbor Laboratory Press

Chapitre V

Biofutur Septembre 2001 , 214 pages à 56 : Dossier spécial « Suppléer la Nature »