

UNIVERSITE DE DROIT, D'ECONOMIE ET DES SCIENCES D'AIX-MARSEILLE  
(AIX-MARSEILLE III)

*N° attribué par la bibliothèque*

00AIX30002

**MÉTHODOLOGIE D'EXTRACTION AUTOMATIQUE  
D'INFORMATION À PARTIR DE LA LITTÉRATURE SCIENTIFIQUE  
EN VUE D'ALIMENTER UN NOUVEAU SYSTÈME D'INFORMATION**

**APPLICATION À LA GÉNÉTIQUE MOLÉCULAIRE  
POUR L'EXTRACTION D'INFORMATION SUR LES INTERACTIONS**

**THESE**

pour obtenir le grade de :

**DOCTEUR EN SCIENCES**

*Discipline : Sciences de l'Information et de la Communication*

présentée et soutenue publiquement

par

**Violaine PILLET**

le 25 janvier 2000

**JURY**

M. P. DESSEN (directeur de recherche) examinateur

M. R. DUCASSE (professeur) rapporteur

M. P. DUMAS (professeur) examinateur

M. B. JACQ (chargé de recherche) codirecteur de thèse

M. L. QUONIAM (professeur) codirecteur de thèse

M. F. RECHENMANN (directeur de recherche) rapporteur

# SOMMAIRE

## I. INTRODUCTION.....8

### PREMIERE PARTIE POSITIONNEMENT DU TRAVAIL

## II. INTRODUCTION BIOLOGIQUE ET LE PROGRAMME CERISE .....13

- A. INTRODUCTION BIOLOGIQUE..... 13
  - 1. *L'information génétique* ..... 13
  - 2. *Organisme modèle choisi : Drosophila melanogaster*..... 14
- B. SYNOPSIS DU PROJET CERISE DANS LE CADRE DU PROGRAMME GENOME - CNRS..... 16

## III. LES SOURCES D'INFORMATION EN BIOLOGIE MOLÉCULAIRE ET GÉNÉTIQUE .....18

- A. TYPES ET STRUCTURES DES SOURCES D'INFORMATION ..... 18
- B. RÉPERTOIRES DE BASES DE DONNÉES ..... 19
- C. BANQUES DE SÉQUENCES..... 20
  - 1. *Banques de séquences nucléiques* ..... 20
  - 2. *Banques de séquences protéiques - Familles de protéines* ..... 21
- D. BASES DE STRUCTURES MOLÉCULAIRES..... 22
- E. BASES DE DONNÉES DE FAMILLES DE MOLÉCULES ..... 23
- F. BASES DE DONNÉES GÉNOMIQUES ..... 23
  - 1. *Bases de données sur les bactéries* ..... 23
  - 2. *Bases de données sur les champignons*..... 24
  - 3. *Bases de données sur les plantes*..... 24
  - 4. *Bases de données sur les animaux* ..... 25
  - 5. *Bases de données sur l'homme* ..... 26
- G. AUTRES BASES DE DONNÉES ..... 26
- H. BASES DE DONNÉES BIBLIOGRAPHIQUES ..... 27
  - 1. *La base de données bibliographique MEDLINE* ..... 27
  - 2. *La base SEQANALREF*..... 28
  - 3. *Autres types de références bibliographiques* ..... 28
- I. CONCLUSIONS SUR LES SOURCES D'INFORMATION EN BIOLOGIE MOLÉCULAIRE ET GÉNÉTIQUE ..... 28

## IV. RECHERCHE ET EXTRACTION AUTOMATIQUE D'INFORMATION .....30

- A. INTRODUCTION ..... 30
- B. LES SYSTÈMES DE RECHERCHE D'INFORMATION..... 30
- C. L'EXTRACTION AUTOMATIQUE D'INFORMATION..... 32
  - 1. *Introduction*..... 32
  - 2. *Définition et enjeux de l'extraction d'information*..... 33
  - 3. *Evaluation des performances d'un processus de recherche ou d'extraction d'information*..... 33
- D. EXTRACTION D'INFORMATION BASÉE SUR DES SYSTÈMES D'ANALYSES LINGUISTIQUES ... 34
- E. EXTRACTION D'INFORMATION BASÉE SUR DE LA STATISTIQUE TEXTUELLE ..... 38
  - 1. *Les méthodes statistiques* ..... 38

2. "Data mining" .....	39
3. "Text mining" (fouille de données).....	39
4. Exemples d'applications dans le domaine biologique.....	40

<p><b>DEUXIEME PARTIE</b>  <b>TRAVAUX EXPÉRIMENTAUX RÉALISÉS :</b>  <b>PRÉPARATION ET ANALYSE STATISTIQUE</b>  <b>D'UN CORPUS SCIENTIFIQUE TEXTUEL</b>  <b>MISE EN PLACE D'UNE MÉTHODOLOGIE D'EXTRACTION D'INTERACTIONS</b></p>
---

<b>V. SOURCE D'INFORMATION TEXTUELLE ET LES OUTILS INFORMATIQUES : LA BASE DE DONNÉES FLYBASE ET LES OUTILS UTILISES.....</b>	<b>41</b>
A. PROBLÉMATIQUE GÉNÉRALE ET SYNOPSIS DU TRAVAIL .....	41
B. LES OUTILS UTILISÉS.....	42
1. <i>Infotrans</i> .....	42
2. <i>Dataview</i> .....	43
3. <i>STATISTICA</i> .....	44
4. <i>Infobank et Idealist</i> .....	44
5. <i>DATALEM</i> .....	44
6. <i>Matrisme</i> .....	44
7. <i>Word, Excel</i> .....	45
C. CHOIX DE LA BASE DE DONNÉES .....	45
D. HISTORIQUE DE LA BASE DE DONNÉES FLYBASE .....	46
E. STRUCTURE DE LA BASE DE DONNÉES.....	46
<b>VI. PRÉPARATION D'UN CORPUS DE RÉFÉRENCES (CORPUS DE REALISATION).....</b>	<b>53</b>
A. LA CHAÎNE DE TRAITEMENTS : EXTRACTION, RESTRUCTURATION .....	53
1. <i>Extraction des champs pertinents</i> .....	53
2. <i>Restructuration du champ "PI" : découpage en phrases</i> .....	54
3. <i>Attribution d'un numéro univoque</i> .....	54
4. <i>Localisation d'entrées non pertinentes</i> .....	56
5. <i>Epuration du corpus</i> .....	56
6. <i>Conclusion des premières étapes de la chaîne de traitements</i> .....	57
7. <i>Dédoublonnage</i> .....	57
8. <i>Conservation des entrées ne concernant que l'espèce melanogaster</i> .....	57
B. LA CHAÎNE DE TRAITEMENTS : RECONNAISSANCE ET BALISAGE DES SYMBOLES DE GÈNES ET DE PROTÉINES .....	58
1. <i>Introduction au problème</i> .....	58
2. <i>Désambiguïsation des noms de gènes</i> .....	58
3. <i>Reformatage (traitement des ponctuations)</i> .....	61
4. <i>Balisage des noms de gènes et de protéines</i> .....	66
C. LA CHAÎNE DE TRAITEMENTS : PARAMÈTRE DE SÉLECTION .....	68
D. SYNOPSIS ET ÉVALUATION À MI-PARCOURS .....	69
E. LA CHAÎNE DE TRAITEMENTS : TRAITEMENT SÉMANTIQUE DES DONNÉES .....	69
1. <i>La normalisation</i> .....	69
2. <i>Mise en place de lexiques terminologiques</i> .....	71
3. <i>Les noms composés</i> .....	72
4. <i>Lemmatisation du corpus</i> .....	74

5. <i>Les mots vides</i> .....	74
6. <i>Lexique de termes spécifiques au domaine</i> .....	76
F. LA CHAÎNE DE TRAITEMENTS : VALIDATION BIOLOGIQUE DU CORPUS.....	78
<b>VII. ANALYSE STATISTIQUE TEXTUELLE DU CORPUS .....</b>	<b>82</b>
A. ANALYSE MULTIDIMENSIONNELLE (AFC) : DÉTERMINATION DU VOCABULAIRE SPÉCIFIQUE .....	82
1. <i>Paramétrage de l'Analyse Factorielle des Correspondances</i> .....	82
2. <i>Résultat de l'Analyse Factorielle des Correspondances</i> .....	84
B. STRATÉGIE DE PRÉDICTION D'INTERACTION .....	87
1. <i>Les requêtes</i> .....	87
2. <i>Les Index de Vraisemblance d'Interaction (IVI)</i> .....	91
<b>VIII. UN NOUVEAU SYSTÈME D'INFORMATION : LA BASE DE DONNÉES FLYNETS-LIST .....</b>	<b>98</b>
A. INTRODUCTION .....	98
B. STRUCTURE ET CONTENU DE FLYNETS-LIST .....	98
C. CONCLUSION .....	100
<b>IX. VISUALISATION CARTOGRAPHIQUE DES DONNÉES SUR LES INTERACTIONS .....</b>	<b>101</b>
A. INTRODUCTION .....	101
B. RÉSEAU GRAPHIQUE D'INTERACTIONS DIRECTES .....	101
1. <i>Réseau d'interactions directes à partir du corpus des 1200 phrases</i> .....	101
2. <i>Réseau d'interactions directes à partir de Flynets-list</i> .....	103
C. CONCLUSION ET PERSPECTIVES .....	104
<b>X. CONCLUSION, DISCUSSION ET PERSPECTIVES .....</b>	<b>106</b>
A. CONCLUSION ET DISCUSSION .....	106
B. PERSPECTIVES À COURT TERME .....	107
1. <i>Mise en place d'un corpus de validation</i> .....	107
2. <i>Validation biologique du corpus</i> .....	107
3. <i>Test de la stratégie sur le corpus de validation</i> .....	108
4. <i>Réajustements de la méthodologie</i> .....	108
C. PERSPECTIVES À LONG TERME .....	109
1. <i>Amélioration de la stratégie</i> .....	109
2. <i>Transposition de la stratégie d'extraction d'information</i> .....	110
<b>XI. LEXIQUE.....</b>	<b>113</b>
<b>XII. BIBLIOGRAPHIE .....</b>	<b>117</b>

## I. INTRODUCTION

Le phénomène informatique bouleverse nos sociétés depuis près de 30 ans. La révolution que constitue la prolifération des données, et surtout, la disponibilité immédiate d'informations nombreuses, variées et de grande qualité, soulève des questions scientifiques, juridiques ou éthiques. Ainsi, l'explosion documentaire, observée depuis la fin des années 1960, aboutit à une croissance quasi-exponentielle du flux d'information.

Ces bouleversements ont conduit les acteurs des sciences de l'information et de la communication "les informaticiens" à développer de nouveaux outils et techniques de recherche, de traitement et de diffusion de l'information. Ceux-ci permettent aux différents utilisateurs (ceux qui recherchent l'information) d'acquérir des documents pertinents et de leur fournir de l'information à forte valeur ajoutée.

D'après Hubert Fondin et Jacques Rouault [FOND99] dans *L'information, l'Arlésienne de l'interdiscipline des sciences de l'information et de la communication* : «La recherche en sciences de l'information doit étendre ou renouveler ses problématiques. Elle doit reconsidérer ses terrains traditionnels et s'intéresser à des domaines nouveaux...». Elle se doit donc d'être interdisciplinaire. C'est pour cette raison que nous proposons, dans ce travail, de mettre les acteurs des sciences de l'information et de la communication au service d'un nouveau domaine riche en potentialités : **la biologie**. Nous proposons un nouveau champ interdisciplinaire qui rapproche les sciences de l'information et la biologie, plus précisément, la biologie moléculaire, la génétique et la bioinformatique.

En effet, il existe une réelle explosion des données dans le domaine de la génétique. Une multitude d'éléments, stockées dans les banques et bases de données, est devenue accessible grâce à la délocalisation que permettent les réseaux de téléinformatique. La généralisation des réseaux tel qu'Internet, du courrier électronique, etc., entraîne, de fait, une surabondance de sources d'information notamment dans le domaine de la biologie. Il existe en biologie une multitude de bases de données factuelles, bibliographiques et textuelles. Nous montrerons, après avoir donné un bref aperçu de la problématique du point de vue biologique, l'apport des Sciences de l'Information et de la Communication dans ce domaine.

Depuis quelques années, le nombre de données produites dans le domaine de la biologie moléculaire et des génomes (souris, homme...) a crû de manière considérable [DESS95]. La quantité et la diversité de données produites par le séquençage de l'ADN, mais aussi par les multiples approches expérimentales de l'étude des génomes conduit le biologiste devant une masse colossale de données, ce qui l'oblige à utiliser des outils spécifiques pour une recherche pertinente des données concernant son domaine.

Les banques de données se sont multipliées, allant du simple catalogue développé dans le cadre d'un travail de recherche isolé, à des projets internationaux. Ainsi, les banques de séquences de biomolécules, comme l'EMBL (European Molecular Biology Laboratory) [RODR96] ou de génomes comme le GDB (Genome DataBase) [FASM96], sont devenues des outils indispensables pour la recherche. Elles constituent la mémoire des données produites dans les laboratoires, rassemblant les résultats publiés dans la littérature scientifique et permettant leur diffusion.

Les programmes de séquençage des génomes fournissent en masse de nouvelles données structurales. La prochaine étape sera d'être capable d'associer à ces données de structures les données fonctionnelles correspondantes. La compréhension de tout processus biologique passe par la caractérisation génétique, moléculaire ou biochimique des différents gènes qui en contrôlent le déroulement. Il s'agit ensuite de mettre en évidence expérimentalement les interactions génétiques et moléculaires existantes entre ces gènes.

A l'avenir, la description et la représentation formelle des interactions, leur assemblage en réseaux régulateurs, la modélisation et la simulation de leur fonctionnement seront des étapes essentielles de la compréhension des processus contrôlés par ces gènes.

Enfin, la comparaison de réseaux de régulation entre organismes différents sera un élément important de l'étude des phénomènes de régulation génique et de conservation fonctionnelle à travers l'évolution.

De nombreuses bases de données stockent les séquences ou les structures des gènes et des protéines, mais il est surprenant de constater qu'aucune base de données n'est consacrée à leur aspect fonctionnel, notamment les interactions moléculaires spécifiques qu'ils établissent entre eux. **Aucune image d'un réseau de régulations et de ses composants n'est actuellement déductible des bases de données disponibles.**

De nombreux travaux décrivent les interactions moléculaires et génétiques intervenant dans différents processus biologiques. Alors qu'il est facile de rechercher dans les bases de données des informations sur un gène ou une protéine donnés, il est en revanche, très difficile de trouver des données sur les interactions entre deux partenaires car celles-ci sont disséminées parmi d'autres types d'information et ne sont pas indexées en tant que telles.

Les biologistes manifestent donc un besoin grandissant d'information sur la nature fonctionnelle des gènes et demandent des documents porteurs de cette information. Mais comment et où rechercher ces données et quelles sont les sources existantes sur ce type d'information ? Ces données existent mais ne sont pas accessibles de manière simple et rapide, et les chercheurs en biologie n'ont ni les moyens ni le temps d'y accéder.

**Les métiers d'intermédiation comme ceux issus des sciences de l'information peuvent rendre possible cette recherche et cette diffusion d'information.**

De par l'interaction entre la biologie et les sciences de l'information pour combler le manque de sources d'information sur les interactions génétiques et moléculaires et pour ensuite représenter, stocker et analyser ces données, est né le consortium CERISE (Consortium d'Etude des Réseaux d'Interactions dans les Systèmes Eucaryotes). C'est un des programmes GENOME du CNRS incluant 5 équipes de recherche. Nous montrerons plus en détail le contenu de ce programme dans le chapitre II.B.

Cette action organisée met en relation plusieurs domaines : l'informatique, les sciences de l'information, la bioinformatique et la biologie. Pour mener à bien ce projet, une réelle concertation et communication entre les différentes équipes sont indispensables. Un des volets de ce programme de recherche a pour but de concevoir, réaliser et maintenir de façon la plus automatique possible, un système d'information (base de données et de connaissances) permettant le recensement et la diffusion des connaissances acquises sur les interactions

génétiques et moléculaires. Ces données sont essentielles à la compréhension des processus biologiques. Cette mémoire informationnelle est réalisable, bien entendu, après un travail préalable de collecte et de traitement automatique des données recueillies. En effet, avant de réaliser un tel système, il est indispensable de sélectionner les sources contenant les données, de les extraire et enfin de les analyser.

Or les données extraites ne sont pas directement exploitables. Une phase de réécriture et de reformulation des données collectées doit permettre de les stocker dans la base sous une forme appropriée. Une vue synthétique des données analysées est obtenue à l'aide de graphes représentant les réseaux formés par les interactions génétiques et moléculaires. Pour réaliser un tel travail technique, l'utilisation d'outils et de concepts en sciences de l'information et de la communication est nécessaire. C'est pourquoi nous nous proposons de mettre en place une méthode capable de recenser automatiquement les données sur les interactions, issues de documents textuels. Celle-ci met en œuvre une stratégie d'extraction d'information faisant intervenir des opérations de filtrage, de remodelage en données, de création d'un système d'information pour accueillir ces données. Elle propose une vue synthétique qui permet de diffuser les connaissances auprès des biologistes demandeurs, par le canal des nouveaux réseaux de communication.

Ce manuscrit se décompose en deux parties et nous allons expliciter brièvement les sujets abordés dans chacune d'elle.

La première partie, comportant 3 chapitres, porte sur le positionnement de notre travail dans le domaine de la biologie et des sciences de l'information.

Le premier chapitre comporte une courte introduction biologique expliquant brièvement les notions de génétique indispensables à une compréhension du sujet. Y est inclus de même un bref synopsis du projet mené par le consortium CERISE expliquant de façon générale le contenu de ce programme et le degré d'intégration de notre travail dans ce large contexte.

Dans le deuxième chapitre sont exposées les différentes sources d'information disponibles dans le domaine de la biologie moléculaire et génétique. Il ne s'agit pas là de répertorier toutes les sources existantes mais de montrer l'hétérogénéité et la multitude de données existantes dans ce domaine. On insistera plus particulièrement sur les bases de données bibliographiques, très riches en données, mais très peu exploitées. En effet, il est très difficile pour un chercheur en biologie d'exploiter manuellement la masse d'information stockée dans de telles bases. Nous montrons qu'il existe des carences dans les sources d'information et précisément dans le domaine d'application de notre travail. Lors de l'initiation du sujet, il n'existait pas à notre connaissance de bases de données dédiées aux interactions génétiques et moléculaires. Pour combler ce manque, le consortium **CERISE** a développé une base de données entièrement dédiée aux interactions - **FlyNets-list** - ainsi qu'une base de connaissance **KNIFE**.

Le chapitre trois aborde les nouveaux systèmes développés spécifiquement pour extraire de façon automatique, à partir de sources de données textuelles, des données spécifiques sur un sujet donné. C'est ce que l'on appelle le domaine de **l'extraction d'information**. Un état de l'art des deux grands courants qui partagent ce domaine est décrit : le premier, basé sur des techniques linguistiques, est explicité de façon générale et illustré de

quelques exemples d'applications ; le deuxième, basé sur des techniques statistiques, est abordé selon le même principe, illustré de même de quelques exemples d'application.

La deuxième partie du rapport, divisée en 5 chapitres, est consacrée à l'exposé de notre travail de recherche.

Le premier chapitre expose la problématique du travail de recherche et la difficulté de la tâche. Suit une définition de la base de données textuelle sur laquelle sont effectués les travaux de recherche. Nous expliquons le choix de la base de données **FlyBase** et l'organisme auquel elle est dédiée. Nous nous attachons ensuite à décrire l'historique et la structure de la base dans le but de montrer où se trouvent les informations recherchées : interactions génétiques et moléculaires. Enfin, une courte description des outils utilisés dans notre travail clôt ce chapitre.

Le deuxième chapitre montre dans le détail les différentes étapes de traitement nécessaires à l'élaboration d'un corpus de références qui servira de base à la réalisation de notre stratégie d'extraction. Il met particulièrement l'accent sur les difficultés rencontrées au cours de ces étapes de traitement. Nous montrons comment extraire les champs pertinents de la base FlyBase, quels types de filtres sont appliqués pour éliminer les données non pertinentes. Nous abordons ensuite une méthode de reconnaissance des entités présentes dans les textes à l'aide d'un dictionnaire qui permet d'effectuer d'autres opérations de filtrage. Enfin, nous exposons les techniques destinées à réduire la diversité des termes présents dans les textes, en les normalisant pour effectuer des analyses sur ceux-ci. Un dernier paragraphe traite de la validation biologique du corpus de texte obtenu pour déterminer réellement où se situent les données sur les interactions.

Le troisième chapitre décrit l'analyse textuelle effectuée sur le corpus. Cette analyse est basée sur des techniques d'analyse statistique de données textuelles et s'applique sur chaque phrase du corpus. Elle permet, dans un premier temps, de déterminer le vocabulaire utilisé pour décrire une interaction. A partir de ces termes spécifiques sont mis en place deux stratégies de prédiction d'interaction. L'une est basée sur les requêtes, l'autre sur un calcul d'Index de Vraisemblance d'Interaction. L'objectif recherché est de pouvoir prédire, avec une marge d'erreur la plus étroite possible, qu'une phrase traite ou non d'interaction. Nous montrerons alors, en fonction des résultats obtenus, la stratégie qui a été retenue.

Le quatrième et le cinquième chapitres exposent la réalisation d'un nouveau système d'information dédié aux interactions. Le quatrième chapitre décrit la nouvelle base de données entièrement dédiée aux interactions : FlyNets-list. La structure de la base et les différentes données accessibles y sont décrites. Cette base est actuellement diffusée via le réseau Internet. Le cinquième chapitre traite de la façon de représenter graphiquement les réseaux formés par ces interactions afin d'en donner une vue synthétique, ce qui, à notre connaissance, n'a jamais été réalisé jusqu'ici. D'un point de vue biologique, il ne s'agit plus de se concentrer sur un gène mais d'obtenir une vision globale des réseaux d'interactions à des fins d'analyse, pour en déduire de nouvelles pistes de recherche.

Nous terminerons ce travail en y montrant que grâce à des techniques statistiques, il est possible d'extraire de façon automatique des données à partir d'une masse importante de données textuelles. Nous montrerons aussi les avantages et les limites de la méthode en la comparant avec les autres techniques exposées dans le troisième chapitre de la première partie.

Enfin, un lexique biologique a été constitué pour aider le lecteur non-biologiste à mieux situer le domaine biologique et à en mieux saisir la problématique. Chaque mot du manuscrit portant un astérisque figure dans le lexique.

## II. INTRODUCTION BIOLOGIQUE ET LE PROGRAMME CERISE

### A. Introduction biologique

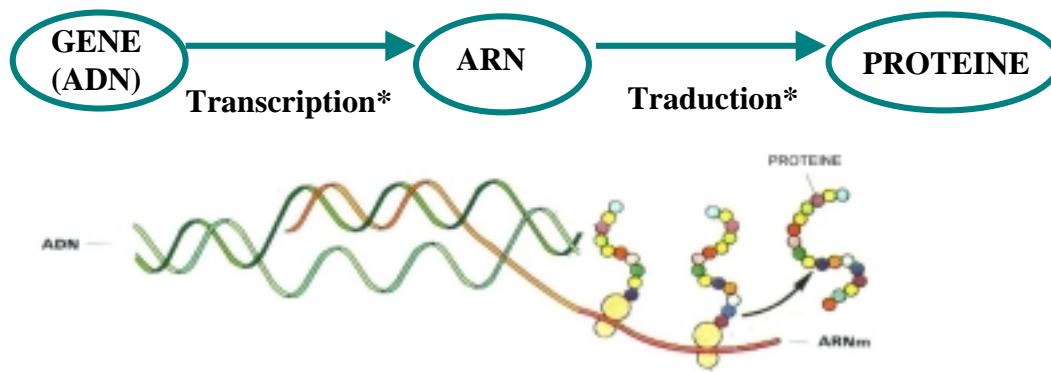
#### 1. L'information génétique

La compréhension des mécanismes qui permettent la formation, le développement\* et le maintien d'un organisme reste l'un des grands défis actuels de la biologie.

Nous savons que tout organisme vivant est composé de **cellules**. Chaque cellule contient un **patrimoine génétique\*** ou encore **génome\***, qui assure la **transmission des caractères\*** spécifiques de génération en génération.

Le génome, qui contient l'ensemble des informations indispensables à la reproduction et au fonctionnement d'un organisme, est constitué d'une macromolécule d'**acide désoxyribonucléique\***, plus communément appelée **ADN\***. Cet ADN contient plusieurs milliers d'unités fonctionnelles, appelées **gènes\*** ou encore **unités de transcription**. Le génome de la drosophile contient 15 000 à 20 000 "unités de transcription" alors qu'on en estime de 50 à plus de 100 000 chez l'homme.

La transmission des caractères spécifiques (ou transmission du patrimoine génétique) s'effectue à partir du génome de chaque cellule (Figure 1). Les gènes (ADN) sont tout d'abord transcrits en **ARN\*** (**acide ribonucléique**). Chaque ARN spécifique est ensuite traduit en **protéine**. Chaque protéine (macromolécule) possède une fonction particulière (enzyme, facteur de transcription...) permettant le fonctionnement et la vie de chaque cellule [LEWI88].



Le fonctionnement correct des cellules implique des **interactions** entre les 3 familles de molécules informationnelles (ADN, ARN, protéine). Nous retiendrons ici la définition suivante pour les interactions moléculaires :

**Il y a une interaction génique entre le gène A et le gène B (ou inversement) si le produit final du gène A (protéine) interagit moléculairement avec le gène B (ADN ou chromatine) ou l'un des produits du gène B (ARN ou protéine).**

Il existe 3 principaux types d'interactions moléculaires :

- **protéine-ADN** (par exemple, un facteur de transcription en se fixant à un site de l'ADN contrôle la transcription d'un gène).
- **protéine-ARN** (par exemple, dans le ribosome, plusieurs protéines reconnaissent des régions spécifiques de l'ARN ribosomique).

- **protéine-protéine** (par exemple, une enzyme protéolytique interagit spécifiquement avec un précurseur hormonal inactif pour le cliver et fabriquer une hormone peptidique active).

Dans la littérature scientifique, 90% des informations publiées sur les interactions moléculaires concernent ces trois types d'interactions. Les interactions génétiques sont vues comme des combinaisons d'interactions directes. Nous nous intéresserons donc à ces données pour essayer dans un premier temps de les recenser et ultérieurement les analyser pour tenter de comprendre la logique de communication qu'utilisent les gènes entre eux.

## 2. *Organisme modèle choisi : Drosophila melanogaster*

L'étude de la fonction du génome passe par l'étude de celui-ci au travers d'organismes modèles. Le choix de ces organismes modèles est justifiée à plusieurs titres :

- parce que pour certains organismes l'étude des gènes et des mutants\* est relativement facile à réaliser,
- parce que certains organismes ont un cycle de reproduction rapide,
- parcequ'ils sont facilement manipulables expérimentalement,
- parce que de nombreuses connaissances ont été accumulées sur certains de ces organismes,
- parce qu'il existe des principes d'homologies\* entre ces organismes modèle et l'homme. L'étude du fonctionnement des gènes chez un organisme modèle peut permettre de mieux comprendre le fonctionnement des gènes chez l'homme grâce aux ressemblances évolutives (homologies) entre les gènes des deux espèces\*.

Pour notre étude, le choix de l'organisme modèle s'est porté sur l'insecte *Drosophila melanogaster*.

La drosophile ("Fruit Fly") est un insecte diptère qui représente un matériel de choix pour la génétique du développement\* car c'est l'organisme eucaryote\* pluricellulaire dont la génétique\* est la mieux connue. Parmi les nombreuses espèces de drosophiles, celle qui est la plus étudiée en laboratoire est *Drosophila melanogaster*.

Le développement de la drosophile comprend plusieurs étapes distinctes. Après la ponte, l'œuf se transforme en un individu adulte en passant par un stade embryonnaire, trois stades larvaires et un stade pupal (Figure 2).

L'étude de la fonction d'un gène au cours du développement de la drosophile est facilitée par l'observation d'embryons\* mutants c'est-à-dire dont un ou plusieurs gènes ne sont pas exprimés ou donnent un produit altéré.

L'expression\* de certains gènes dans ces mutants permet d'avoir une idée des interactions entre gènes ainsi que leur ordre d'action. L'étude approfondie de ces interactions géniques permet de mieux comprendre les processus développementaux de la drosophile. Pour ce faire, il est indispensable de rassembler le maximum de données relatives aux gènes régulateurs et à leur expression spatio-temporelle.

Actuellement, une centaine de gènes régulateurs ont été identifiés. La prise en compte simultanée des données moléculaires, génétiques et de celles concernant leur expression dépasse les capacités du cerveau humain. L'accès à cet ensemble de connaissances peut être facilité par la création d'une base de données informatisée. Ce type de base permet l'organisation et la gestion des données ainsi que leur exploitation.

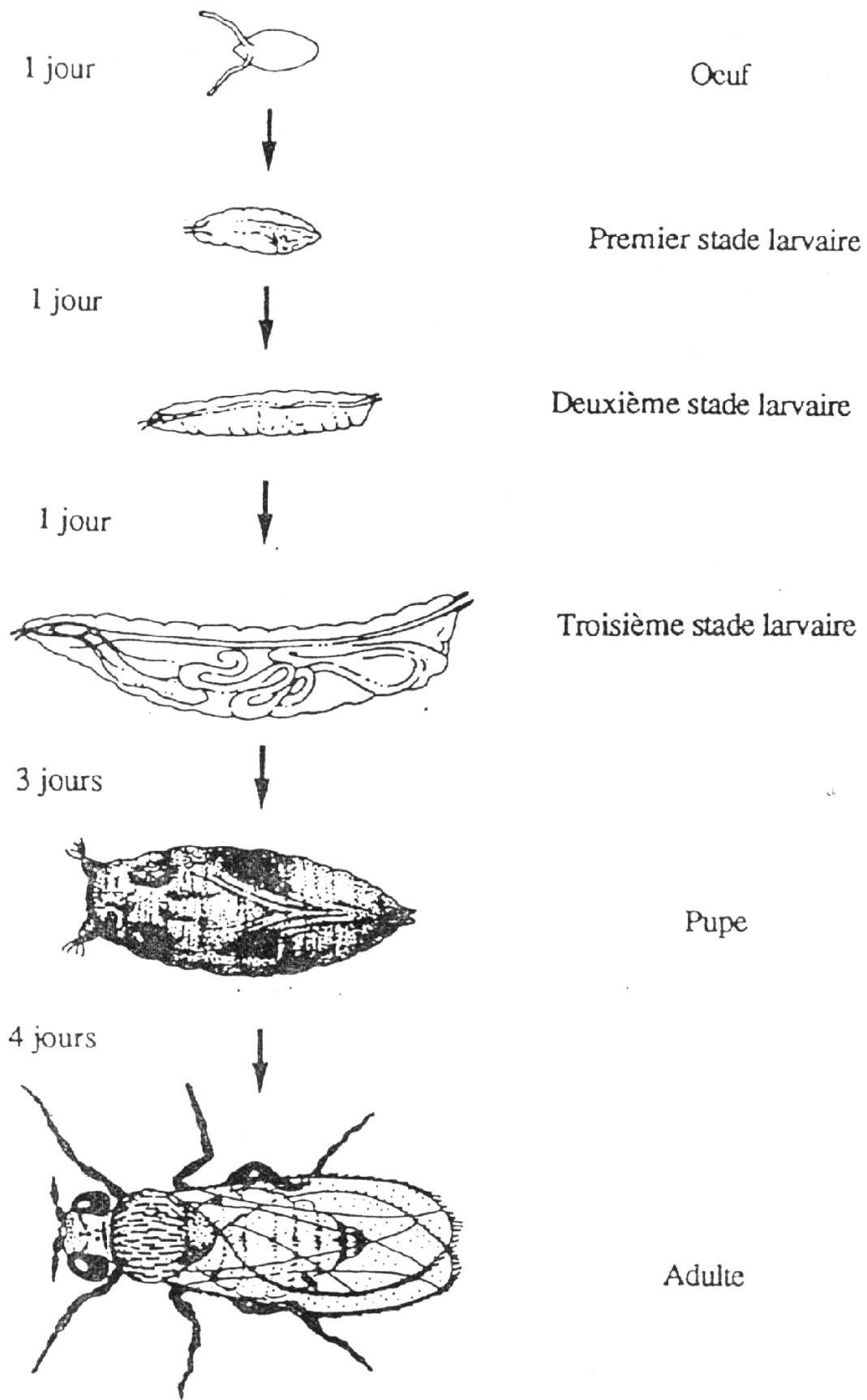


Figure 2 : diagramme du cycle de vie de *Drosophila melanogaster*

## **B. Synopsis du projet CERISE dans le cadre du programme GENOME - CNRS**

Le programme CERISE est une approche bioinformatique **pluridisciplinaire** de la fonction et de l'évolution des génomes à travers l'étude des réseaux régulateurs géniques. Il se situe donc clairement dans le cadre de ce que l'on appelle couramment maintenant les projets d'après-génome.

Ce consortium met en jeu 5 équipes de recherche de spécialités différentes : deux laboratoires en biologie, un en biologie-bioinformatique, un en informatique et un en sciences de l'information.

Il vise à saisir, représenter formellement et exploiter (analyser, comparer, simuler) des données sur les **interactions génétiques et moléculaires** entre gènes, ARN et protéines et sur les réseaux de régulations fonctionnelles qu'elles constituent. L'exploitation de ces données qui n'avaient jusqu'ici fait l'objet d'**aucune représentation** informatique se fera dans 3 directions : l'analyse des graphes que représentent ces réseaux, la comparaison de réseaux, au sein d'un même organisme et entre organismes différents, et la simulation d'aspects spécifiques du fonctionnement de ces réseaux.

Deux modèles expérimentaux eucaryotes pluricellulaires seront utilisés comme source de données et objets d'analyse : la mouche *Drosophila melanogaster* et la souris, avec dans les deux cas une attention particulière, mais non exclusive, apportée aux régulations\* géniques développementales. Grâce à différentes approches **complémentaires**, utilisant d'une part des technologies récentes de la représentation des connaissances par objets et de la simulation qualitative et d'autre part celles apparues en sciences de l'information, il s'agira d'organiser ces données sur les interactions et leurs réseaux au sein de deux nouvelles bases : la base de données **FlyNets-list** et la base de connaissance KNIFE (Knowledge of Networks of Interactions in the Fly and other Eukaryotes).

Ce consortium établira des cartes d'interactions fonctionnelles génomiques basées sur les 3 types essentiels d'**interactions moléculaires** (protéines-ADN, protéines-ARN et protéines-protéines) et sur les **interactions génétiques** indirectes, vues comme des combinaisons d'interactions directes. Le volet biologie évolutive de ce programme permettra pour la première fois d'obtenir des données d'évolution fonctionnelle sur la conservation des réseaux d'interaction, qui sont un complément indispensable aux données évolutives structurales entre séquences homologues\*. Ce dernier aspect impliquera la réalisation de la base HOMOGEN qui précisera les relations d'homologie\*, orthologie\* et paralogie\* entre tous les gènes existants.

Le programme général de ce projet est subdivisé en quatre volets, relativement indépendants dans leur réalisation initiale, mais fortement complémentaires dans leurs résultats, qui sont :

<p><b>acquisition automatique de données sur les interactions à partir de textes électroniques.</b> Le travail présenté dans ce manuscrit résulte d'une collaboration entre la biologie et les sciences de l'information et s'intéresse dans ce projet surtout à cet aspect-là. Nous verrons donc en détail dans ce manuscrit comment il est possible, avec des outils d'analyse statistique de données textuelles, d'extraire des données sur les interactions à partir d'un gros volume de données textuelles.</p>
--

- **conception et réalisation d'un système d'information** : stockage des données extraites dans une base de données - **FlyNets-list** - dédiées aux interactions impliquant les gènes de la drosophile et réalisation d'une base de connaissance - **KNIFE** - sur ces interactions. Puis représentation sous forme de graphes des réseaux que forment ces interactions. Nous montrerons aussi un peu plus en détail en fin du manuscrit, cette partie du projet.

- l'analyse de ces réseaux et la simulation de certains aspects du fonctionnement des réseaux.

- l'étude de l'évolution biologique des réseaux régulateurs eucaryotes.

Chaque volet est pris en charge plus spécifiquement par une des équipes, tout en favorisant au maximum les interactions avec les autres partenaires.

Ce programme propose donc d'aborder une même question biologique, celle de l'étude conceptuelle des réseaux d'interactions moléculaires et génétiques, par une approche **pluridisciplinaire** faisant intervenir une convergence originale entre les sciences de l'information, l'informatique des bases de connaissances et de la simulation, la bioinformatique des comparaisons de séquences\*, la biologie du développement et enfin l'évolution moléculaire en établissant des rapports originaux entre évolution structurale et fonctionnelle.

Parmi les nombreuses retombées qui peuvent être prévues pour ce projet, nous retiendrons la création future, grâce à la syntenie souris-homme, de la première carte des **interactions fonctionnelles du génome humain**.