Chapitre 1

Introduction

"Would you tell me, please, which way I ought to go from here?"
"That depends a great deal on where you want to go," said the Cat.
"I don't much care where -" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"- so long as I get somewhere," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Alice in Wonderland. Lewis Carroll.

Imaginez que l'on place devant vous une image — par exemple, la photographie d'une scène de la vie quotidienne un jour de printemps autour du bassin du Jardin du Luxembourg — et que l'on vous demande de dire si cette image contient un petit chien, et si oui, où se situe l'animal. Supposez que l'on vous fournisse un dessin du chien afin de vous guider, et que ce dessin soit identique à celui du petit chien illustré dans l'image (s'il s'y trouve). Dans ce cas, il suffit de glisser un décalque de ce dessin sur toute l'étendue de l'image pour être à même de répondre à la question que l'on vous a posée. Si, par contre, le chien du dessin que l'on vous a donné apparaît de façon différente de celui que vous devez essayer de localiser (par exemple, debout de face, alors que dans l'image le chien est couché), il vous faudra repérer d'abord quelles sont les caractéristiques essentielles de l'animal (sa forme, son apparence générale) afin que le dessin puisse vous être utile dans votre examen de l'image. Enfin, si rien ne vous est fourni comme aide dans votre recherche, il vous faudra recréer un dessin mental de ce qui fait d'un chien un chien (et non un chat ou une souris) et c'est ce 'dessin' qui doit alors vous guider. Notez que j'ai dit recréer car si vous n'avez jamais vu un chien dans votre vie, ni lu une description de l'animal, vous serez impuissant à répondre à la question, même si l'image contient vraiment un chien. Dans tous les cas, il vous faut donc avoir une idée au moins *a priori*, concrète ou abstraite, de ce que vous cherchez.

Imaginez maintenant que l'on pose devant vous non plus une image, mais toute une série de scènes — toujours de notre bassin du parc — dont chacune capte cette fois un instant différent de la vie, par exemple trente photographies, chacune représentant un jour du mois de juin à la même heure du matin. Ce que l'on vous demande à présent est de dire si vous voyez quelque chose d'inhabituel dans ces cinquantes images. La première question qui vous vient à l'esprit est sans doute, qu'est-ce que l'inhabituel? Cela peut être une singularité, quelque chose qui n'est aperçue qu'une fois. Parmi toutes ces images qui fourmillent de détails, beaucoup d'objets sont singuliers, mais, par exemple, ce premier lundi de juin, toutes les chaises étaient renversées par terre et cela attire votre attention. Si vous avez une idée des choses qui sont 'normales' dans la vie, vous vous dites que cela est certainement inhabituel. Bien sûr, il vous faut avoir l'idée d'abord.

L'inhabituel parmi tous ces objets et personnages qui changent d'un jour à l'autre peut cependant être considéré non comme quelque chose de rare — un évènement comme celui des chaises renversées — mais plutôt comme quelque chose qui réapparaît de manière régulière. Dans toutes ces images vous observez par exemple un petit garçon qui regarde tout joyeux un voilier en papier glisser à la surface de l'eau. C'est toujours le même petit garçon, le même voilier tout blanc, la même expression de bonheur.

Identifier ce second type d''inhabituel' est plus facile qu'identifier le premier. Ce n'est pas aussi facile que de localiser un chien (dont on possède un décalque ou une simple description) parce que l'objet qui apparaît souvent n'est pas connu à l'avance. Dans ces images, c'est un petit garçon et son voilier, ailleurs ce sera autre chose. Une comparaison des images, surtout si l'on dispose de beaucoup d'images (et d'une bonne mémoire), facilite cependant la recherche. La finesse de la réponse que nous fournissons à la question : dites-moi ce qui est inhabituel parmi les images, dépend toutefois encore beaucoup de pouvoir avoir une idée *a priori* de ce qui est important parmi tous les détails que ces images présentent. Par exemple, les chaises apparaissent tous les jours, mais vous ne serez sans doute pas tenté d'avoir votre attention attirée par cela car vous savez que, dans toutes les photographies de parcs, n'importe où dans le monde, des chaises similaires seront présentes. Mais il faut que vous sachiez cela pour pouvoir tracer la frontière entre ce qui est juste inhabituellement fréquent, et ce qui est beaucoup trop fréquent et donc probablement sans importance. Parfois également, ce qui va compter n'est pas un objet qui se répète, mais plusieurs objets qui sont vus souvent ensemble. Ainsi, vous remarquez que tous les mardis et mercredis, le petit garçon apparaît avec une écharpe rouge autour du cou. Or ces jours-là justement — ces jours-là uniquement — vous observez également une vieille dame assise sur une des chaises du parc, proche de l'enfant. C'est sa grand-mère, c'est elle aussi qui explique l'écharpe : elle a trop peur qu'il prenne froid. Pour vous, c'est peut être cela qui est inhabituel, et intéressant, dans ces images : le fait qu'écharpe rouge et vieille dame sont présents toujours ensemble (association d'objets) et que la vieille dame est assise près de l'enfant (lieu de la répétition). Bien sûr cela aurait pu être une simple coïncidence : à l'heure où l'enfant joue dans le parc, se trouve présent également un vieux monsieur avec une casquette verte. Les deux ne vont pourtant pas ensemble, mais vous n'avez aucun moyen de le savoir. Ce que vous êtes capable de dire c'est uniquement : cet objet, ou personnage, apparaît tant de fois dans toutes ces images. Par contre, au contraire de ce qui se passe dans la recherche d'un objet spécifique qui doit être connu *a priori*, ici vous pouvez identifier des choses surprenantes, des choses que vous ne connaissiez pas à l'avance : il suffit qu'elles se répètent suffisamment souvent pour attirer votre regard, et que vous soyez sans parti pris. Car en effet, si vous n'aimez pas les oiseaux, vous ne dépenserez sans doute pas de temps à les regarder et vous ne remarquerez pas le petit rouge-gorge qui tous les jours vient se poser au bord du bassin.

Les 'images' qui vont être le sujet de ce travail, celles que nous allons étudier et analyser au moyen d'algorithmes, représentent des macromolécules biologiques. Celles qui vont nous intéresser sont de deux types : les acides nucléiques (ADN et ARN) et les protéines. Les premières peuvent être vues comme les agents transporteurs et transmetteurs du matériel génétique nécessaire au maintien de la vie, et les secondes comme les agents exécuteurs des instructions contenues dans ce matériel. Nous avons choisi pour l'instant de ne les regarder qu'à travers une fenêtre qui nous permet de les voir comme des chaînes de caractères sur un certain alphabet. Cette fenêtre est 'naturelle' car bien que ces macromolécules soient en fait des objets tridimensionnels, ce sont également des polymères linéaires composés d'éléments plus simples — des monomères appelés nucléotides (de quatre types différents) dans le cas de l'ADN et de l'ARN, et des acides aminés (au nombre de vingt) dans le cas des protéines. Cette vision unidimensionnelle des macromolécules ne nous empêche pas de les voir également comme des formes dans l'espace, mais ce sont alors des formes contraintes par l'ordre de succession des monomères le long de la chaîne et codées sur un alphabet d'angles discrétisés (ceux que réalise la chaîne en se repliant sur elle-même). Dans tous les cas, ce sont les symboles associés à chacun de ces monomères (en tant que tels ou en tant qu'objets d'une chaîne 3D) qui vont composer notre alphabet. L'analyse que nous allons faire des macromolécules est donc une analyse lexicale.

La recherche de ce qui est commun à un ensemble d'objets présente une analogie avec ce dont nous avons discuté auparavant à propos de la recherche d'objets communs à diverses images d'un parc. En effet, si nos 'images' ne vont pas représenter le même objet, suggérant que nous aurions dû parler d'images de parcs différents plutôt que de celles d'un même parc à divers moments de la vie, le temps va souvent constituer un lien entre les objets qui vont nous intéresser ici. Ce temps est celui de l'évolution et les macromolécules que nous allons vouloir comparer peuvent être vues comme des 'copies', ou des copies de copies différentes d'une même macromolécule ou d'une même forme ancestrale, chaque copie étant le résultat d'une suite particulière de transformations et de mutations effectuées sur des copies antérieures. Certaines parties de ces copies cependant vont 'changer moins' que d'autres, et nous savons que ces parties sont justement celles qui ont le plus de chance d'être associées, soit à une activité de la chaîne polymérique, soit à un élément structural qui sert de charpente à la structure générale que la macromolécule adopte dans l'espace. Ce sont essentiellement ces parties qui nous intéressent et dans notre analyse lexicale des chaînes de symboles codant pour ces molécules, nous allons donc vouloir identifier des mots dans ces chaînes similaires entre eux d'une certaine façon.

Une première difficulté à laquelle nous allons être confrontés dans cette tâche est alors de définir plus précisément ce que ces termes vagues de 'changent moins' et 'similaires entre eux d'une certaine façon' veulent dire. Une partie essentielle de ce travail est ainsi composée par une exploration de la notion de ressemblance entre macromolécules biologiques. Les monomères sont le point de départ de cette exploration et leur ressemblance est essentiellement une fonction des propriétés physico-chimiques que certains d'entre eux possèdent en commun. La relation cependant n'est pas simple car le rôle exercé par ces monomères dans une macromolécule dépend de plusieurs autres facteurs que leur seule identité, comme leur position dans la chaîne et les liaisons chimiques qu'ils réussissent à établir avec d'autres monomères. Cette influence du contexte ne peut être captée pour l'instant que de manière indirecte et locale. Aussi d'une façon générale, nous devons considérer que les mesures de similarité que nous pouvons établir sont indépendantes du contexte. Ces mesures sont soit discrètes (deux monomères sont considérés comme semblables ou non), soit représentent une valeur numérique. C'est sur la base de ces mesures de similarité entre monomères que nous définissons alors celles entre mots dans les chaînes.

Deux notions sont fondamentales dans les définitions de ressemblance que nous établissons : l'une est celle d'une comparaison multiple, l'autre celle du modèle. L'idée de la première est que comparer un grand nombre d'objets simultanément nous permet d'être beaucoup plus sensibles, c'est-à-dire, de détecter de plus faibles similarités entre ces objets. Le second concept est ce qui permet de réaliser cette comparaison simultanée de manière efficace. Ce n'est pas un concept nouveau, nous n'avons fait que l'étendre et l'adapter à nos besoins, et le rendre le plus souple possible. Cet objet, qui est externe aux chaînes que l'on compare, peut être soit un mot sur le même alphabet, soit un ensemble de mots, c'est-à-dire, un produit cartésien de sous-ensembles des symboles composant l'alphabet des chaînes. La façon dont nous allons nous en servir pour notre propos consiste à dire que les mots d'un ensemble ne sont similaires entre eux que s'il existe un modèle auquel tous ressemblent. Réciproquement, un modèle n'est considéré comme intéressant ('inhabituel') que s'il existe au moins un certain nombre de mots appartenant à des chaînes différentes qui ressemblent à ce modèle (nous disons de ces mots qu'ils sont des occurrences du modèle dans les chaînes et du modèle qu'il est présent dans les chaînes correspondantes). Notre recherche de mots similaires communs se résume alors à une recherche efficace de ces modèles présents dans au moins un certain nombre des chaînes de l'ensemble que nous comparons. Comme pour les images du parc, nous repérons les modèles individuellement. Si deux ou plus de ces modèles sont identifiés par rapport à un même ensemble de chaînes, nous ne sommes pas capables pour le moment de dire si leur association a un sens ou est fortuite.

Notre problème de fournir une définition de la ressemblance entre mots devient donc à présent celui de définir la ressemblance entre des mots et un modèle. Nous établissons en fait plusieurs de ces définitions dans ce travail, pour des raisons à la fois historiques et pragmatiques. La raison historique est que ces définitions ont, elles aussi, évoluées et se sont affinées au fil du temps. Certaines des définitions initiales, par exemple celle ayant pour base une distance d'édition entre mots, se sont retrouvées contenues dans celles établies plus tard, d'autres sont demeurées à part et répondent à des besoins différents. En effet, il n'y a sans doute pas une seule définition possible de la ressemblance, et chacune est adaptée à un problème particulier. Par pragmatisme donc, nous en maintenons plusieurs. Un souci est cependant demeuré constant par rapport à toutes : que ces définitions soient mathématiquement précises. L'idée est de donner au biologiste (la personne intéressée par ce travail) les moyens d'interpréter avec le moins d'ambiguité possible les résultats obtenus par une application de ces définitions. Pour cela, il faut également que les algorithmes qui recherchent les modèles possèdent certaines caractéristiques.

La principale de ces caractéristiques est l'exhaustivité de la recherche qu'ils effectuent. Une autre partie essentielle de ce travail a ainsi porté sur l'élaboration d'algorithmes combinatoires qui explorent tout l'espace des chaînes que l'on compare et qui fournissent en résultat une liste de tous les modèles, d'une longueur donnée ou de longueur maximale, présents, selon une des définitions de ressemblance entre modèle et mots, dans au moins un certain nombre des chaînes. Chacune de ces définitions a donné lieu à un algorithme spécifique. Le principe de base de ces algorithmes est cependant le même et s'appuie sur une formule de récurrence. Celle-ci peut être vue comme une recette de construction des modèles d'une certaine longueur (i.e. d'identification de ses occurrences) à partir de ceux de longueur plus petite, construction qui cesse pour un modèle donné dès que celui-ci ne vérifie plus la condition de présence sur un nombre minimal des chaînes de départ. Malgré cette condition d'arrêt, la nature combinatoire de la construction pose bien sûr un problème important de complexité algorithmique. Nous proposons ici des algorithmes qui sont tous linéaires dans la longueur totale des chaînes, et éventuellement exponentiels uniquement dans le degré de souplesse autorisé dans la ressemblance entre mots et modèles. Ce facteur exponentiel est intrinsèque au problème et représente toujours une situation de pire cas. Observons au passage que, bien que nous n'ayons parlé jusqu'à présent que de l'analyse d'un ensemble de plusieurs objets, il est clair que n'importe lequel de ces algorithmes peut être utilisé afin de rechercher des mots qui apparaissent répétés dans une seule chaîne.

Il nous a semblé important que ce travail présente également une synthèse de toute une longue série de travaux effectués avant nous sur ce même thème, ou sur des thèmes proches. Il y a deux raisons à cela. La première est que nos définitions et algorithmes se sont souvent inspirés de ceux élaborés par d'autres et nous voulions replacer les notres dans leur contexte. La seconde raison est que nous pensons qu'une étude aussi complète que possible de tout ce qui a été fait dans un domaine, que ce soit au niveau des concepts que de celui des algorithmes, représente une base solide et nécessaire avant de partir vers des chemins complètement nouveaux. Notre contribution à ce domaine a été à la fois :

1. de rendre beaucoup plus souples les définitions de ressemblance entre mots dans les chaînes;

2. de formaliser de manière plus précise toutes ces définitions;

3. enfin, de conjuguer comparaison multiple de chaînes macromoléculaires avec une exploration combinatoire exacte de l'espace de ces chaînes d'une façon qui est relativement efficace en pratique.

Ce travail représente donc à la fois une continuation et une innovation par rapport à ce qui existait déjà.

Le texte est divisé en six chapitres. Le premier offre une introduction générale au problème biologique. Bien que nullement complète, elle essaye de donner un aperçu assez approfondi de tout ce qui peut motiver la recherche que nous réalisons ici. La première partie de cette introduction décrit ainsi les mécanismes biochimiques fondamentaux tandis que la seconde est plus tournée vers les motivations de ce travail. Le chapitre 3 poursuit cette étude du problème biologique puisque nous y réalisons une exploration de la notion de ressemblance, entre monomères d'abord, puis entre chaînes et mots dans ces chaînes. C'est dans les deux dernières sections de ce chapitre que nous introduisons les diverses définitions de ressemblance. Nous y présentons non seulement les notres mais également celles qui ont été proposées avant nous et sur lesquelles nous nous sommes souvent appuyés. À la base de toutes ces définitions se trouve en fait un même concept, qui est celui d'un alignement. Ce dernier peut être considéré comme une mise en correspondance de divers objets en vue de les comparer. Là où nos définitions diffèrent de celles établies antérieurement est que les notres utilisent, comme nous l'avons vu, la notion d'un modèle. Les mots dans les chaînes sont donc comparés de manière indirecte, alors que dans les autres cas, cette comparaison est directe. Les méthodes de comparaison directe d'objets, chaînes ou mots dans les chaînes, sont présentées dans le chapitre 4. C'est là que nous réalisons une synthèse des travaux effectués dans ce domaine avant nous. Puis, dans le chapitre suivant, nous introduisons nos propres méthodes indirectes. Ce chapitre contient ainsi une description détaillée des divers algorithmes que nous avons élaborés, chacun permettant d'identifier les modèles présents au moins un certain nombre de fois dans les chaînes d'un ensemble. Ces deux chapitres sont plus purement informatiques, et ce n'est que dans le cinquième que nous revenons à la biologie puisque ce chapitre est consacré à des illustrations. Nous y présentons divers résultats obtenus avec nos méthodes sur un certain nombre d'exemples. La plupart de ceux-ci ont été choisis dans la littérature et sont considérés comme 'très difficiles'. Notre intention en testant sur eux nos algorithmes était de mesurer leur efficacité, robustesse et intérêt. Ce chapitre comporte ainsi une discussion de ce que ces algorithmes peuvent faire, mais également de ce qu'ils ne sont pas capables pour l'instant de réaliser, ainsi que les éventuelles difficultés liées à leur utilisation. Cette discussion ouvre la voie vers les perspectives offertes par ce travail. Notre sentiment est que celui-ci doit se transformer et aller beaucoup plus loin. En particulier, l'impression, mais elle est très personnelle, est que certaines notions qui sont depuis longtemps acceptées dans le domaine, et que nous-mêmes avons adoptées ici (comparaison multiple, distance d'édition entre chaînes etc) ne sont pas encore totalement mûres ou parfaitement réfléchies. Certaines sont peut être même inadéquates. Le chapitre 7 est consacré alors à une remise en cause de quelques-unes de ces notions. Nous suggérons également quelques voies nouvelles possibles.

Observons pour terminer que ce texte long a été conçu pour des personnes de formations et de goûts divers, et organisé en ce sens dans la mesure du possible. Ainsi, une personne intéressée uniquement par l'aspect biologique du problème consultera essentiellement le chapitre 2, le chapitre 3, en particulier les sections 1, 2 et 3 contenant les notions générales, ainsi que le chapitre 6. Le lecteur concerné par la partie informatique peut ne lire que les deux dernières sections du chapitre 3 (Définitions de ressemblance), le chapitre 4 et le chapitre 5. Ceux qui veulent avoir un aperçu général de nos algorithmes, mais ne veulent pas entrer dans le détail de leur fonctionnement, peuvent se contenter de lire la première moitié du chapitre 5. Enfin, le chapitre 7 devrait, nous l'espérons, intéresser tout le monde.