SYNTHESE

Machines spécialisées pour la comparaison de séquences biologiques

Louis Audoire, Jean Jacques Codani †
Dominique Lavenier, Patrice Quinton, ‡

‡ *IRISA*
Campus de Beaulieu
35042 Rennes Cedex

† *INRIA Rocquencourt*
Domaine de Voluceau
78153 Le Chesnay Cedex

ce travail est supporté par le GREG : Groupement de Recherches et d'Etudes sur les Génomes

RÉSUMÉ. *Cet article présente plusieurs machines parallèles spécialisées pour la comparaison de séquences biologiques. Ce sont des machines principalement bâties autour d'un réseau linéaire de processeurs. Leurs performances dépassent de plusieurs ordres de grandeur celles des machines programmables, permettant ainsi de faire face à l'accroissement extrêmement rapide des banques de séquences.*

ABSTRACT. *This article presents several machines dedicated to biological sequence comparison. These machines are parallel machines based primarily on linear arrays. Their performance is several orders-of-magnitude better than that of programmable machines, allowing them to face the challenge of the extremely fast growth of biological-sequence databases.*

MOTS-CLÉS : *comparaison de séquences, machines parallèles, réseau systolique.*
KEY WORDS : *sequence comparison, parallel machines, systolic array.*

1. Introduction

Le nombre des séquences répertoriées dans les bases de données biologiques s'accroit d'environ 50 pour cent chaque année.

En dépit de l'augmentation constante des performances des calculateurs, l'analyse de ces données sur des machines classiques – processeurs séquentiels du commerce – devient de plus en plus longue. Il faut donc rechercher des

solutions informatiques nouvelles. Trois approches sont possibles : utiliser des *réseaux de calculateurs*, des *calculateurs massivement parallèles*, ou réaliser des *machines spécialisées*.

Répartir un programme de traitement de séquences biologiques sur un réseau de stations de travail est une méthode intéressante. Cette approche a, du reste, été testée par un des auteurs [COD 91] [BEL 92] dans le cadre de l'établissement de la cartographie physique de l'ensemble du génome humain. Cette méthode est cependant limitée par le nombre de stations de travail accessibles et disponibles – en pratique, quelques dizaines au plus.

L'usage de *calculateurs massivement parallèles* permet de réduire fortement la durée des comparaisons. Certains logiciels, tels que *Blaze* ou *MPsrch*, sont déjà disponibles sur des calculateurs massivement parallèles de type MasPar [NIC 90]. Ces logiciels sont compétitifs par rapport à des programmes similaires – Fasta [LIP 85] [PEA 88] ou Blast [ALT 90] – optimisés pour des machines séquentielles : à temps de calcul comparables, les résultats des machines parallèles semblent de meilleure qualité.

La confrontation d'une séquence avec une banque de données qui en comporte plusieurs dizaines de milliers prend quelques secondes sur un calculateur massivement parallèle. Un tel temps de réponse est satisfaisant pour la plupart des applications actuelles, dont le but est de comparer quelques séquences issues du séquençage aux différentes banques existantes. Dans les années à venir, on peut estimer que l'accroissement du nombre de séquences dans les banques sera vraisemblablement compensé par l'augmentation des performances des machines, et que le temps de réponse restera approximativement constant.

Mais l'analyse ou la structuration des banques de données amènent à effectuer des traitements intra ou inter banques [SON 94], dont la durée croit avec le carré de la taille des banques. A titre d'exemple, l'extraction de sous-séquences homologues dans une banque de 20 000 séquences protéiques (Swiss-Prot 21, par exemple) demande approximativement quatre jours sur une machine MasPar MP-1 à 1024 processeurs, en appliquant une méthode très rigoureuse.

En supposant que le temps de comparaison d'une séquence avec une banque est constant, le temps de calcul nécessaire à la confrontation intra ou inter banques est proportionnel au nombre des séquences des banques. Ainsi, l'analyse d'une banque de 200 000 séquences protéiques (taille estimée de la banque Pir vers 1996 ou 1997) sur une machine dix fois plus rapide – ou ayant dix fois plus de processeurs – durerait 40 jours, ce qui est évidemment irréaliste.

On peut améliorer les performances d'un traitement informatique d'un ordre de grandeur en faisant appel à des *architectures spécialisées*, parallèles ou non. En spécialisant les processeurs, on peut gagner aisément un facteur 10 à 100 sur la vitesse des traitements (voir [LEM 93], qui décrit l'utilisation d'un processeur spécialisé pour la recherche de motifs).

En outre, les machines parallèles spécialisées sont bien moins encombrantes que les machines programmables, et il est donc possible de gagner aussi un ordre de grandeur sur les performances en augmentant le nombre de processeurs.

Cet article présente un panorama de quelques machines – ou des projets

de machines – parallèles spécialisées destinées à accélérer la comparaison des séquences biologiques. Dans le paragraphe 2, nous précisons d'abord la gamme d'algorithmes visés. Le paragraphe suivant présente la structure (commune) de ces machines, toutes construites sur le modèle *systolique*. Ces machines sont décrites dans le paragraphe 4, puis comparées dans le paragraphe 5.

2. Les algorithmes de comparaison de séquences

On entend par algorithme de comparaison de séquences les algorithmes qui permettent de résoudre – entre autre – les problèmes suivants :

- recherche de motifs : on dispose d'un jeu de motifs (courtes séquences particulières) et on détermine s'ils appartiennent à une ou plusieurs séquences ;
- recherche de profils : on établit un squelette de séquence et on recherche celles qui s'en rapprochent le plus ;
- comparaison de séquences : on évalue le degré de ressemblance entre deux ou plusieurs séquences ;
- recherche de segments homologues : on localise des régions qui présentent des ressemblances.

Pour chacun de ces problèmes, il existe des méthodes plus ou moins strictes pour évaluer les ressemblances. Elles vont d'une correspondance exacte à une correspondance approximative. Ces dernières sont, en général, très coûteuses en calculs ; elles font appel à des méthodes de programmation dynamique dont la complexité est proportionnelle au carré des tailles des séquences.

C'est dans cette dernière catégorie d'algorithmes que les machines spécialisées sont exploitées. En effet, ces algorithmes présentent certaines caractéristiques – comme la régularité – qui se prètent extrêmement bien à une mise en oeuvre sur des machines parallèles.

A titre d'exemple, nous présentons un algorithme très utilisé et représentatif des complexités mises en jeu ; il a été proposé par Smith et Waterman [SMI 81] pour déterminer des sous-séquences homologues entre deux séquences biologiques. Cet algorithme permet d'identifier des portions de séquences similaires en prenant en compte les erreurs de substitution et les erreurs d'insertion/omission (ou gap) multiples.

A partir de 2 séquences $S1$ et $S2$ de longueur respevtives $l1$ et $l2$, l'algorithme calcule une matrice de valeurs H de taille $l1 \times l2$; Chaque valeur $H(i,j)$ représente une vraissemblance locale et est déterminée par la relation de récurrence suivante :

$$H(i,j) = Max \begin{cases} 0 \\ Max_{1 \leq k \leq i}(H(i-k,j) - g_k) \\ Max_{1 \leq l \leq j}(H(i,j-l) - g_l) \\ H(i-1,j-1) + sub(S1_i, S2_j) \end{cases} \quad (1)$$

avec les initialisations : $H(i,0) = 0 \quad (0 \leq i \leq l1)$ et $H(0,j) = 0 \quad (0 \leq j \leq l2)$

$sub(S1_i, S2_j)$ représente le coût de substitution du caractère $S1_i$ par le caractère $S2_j$. g_k représente le coût d'un gap de k caractères ; il est déterminé par : $g_k = \alpha + \beta \times (k-1)$, α étant le coût du premier gap et β le coût des suivants.

Lorsque la matrice a été calculée, l'examen de ces composantes permet de détecter les endroits où des similarités locales apparaissent. Une procédure de *backtrack* permet ensuite d'identifier complètement les sous-séquences à partir du maximum local trouvé.

En fait, la complexité de cet algorithme – et plus généralement la complexité des algorithmes du même type – est surtout fonction de la première étape. En effet, si on a pris soin, au cours de cette étape, de mémoriser les maxima locaux (s'ils existent et s'ils sont significatifs), la phase suivante qui récupère les sous-séquences similaires a un temps de calcul très petit (voire négligeable) devant l'établissement de la matrice. Dans la pratique, lorsqu'une séquence est comparée à une base de données, seules quelques séquences présentent des homologies locales potentielles. La deuxième phase est alors extrêmement réduite.

Les machines qui sont présentées dans les paragraphes suivants accélèrent toutes ce type d'algorithme. Suivant la relation de récurrence considérée des problèmes plus ou moins complexes peuvent être résolus. L'unité qui permet de mesurer (et de comparer) la complexité des algorithmes est appelée *calcul matriciel élémentaire* ou CME. Elle correpond au calcul nécessaire pour déterminer une composante de la matrice (typiquement quelques opérations arithmétiques).

3. Machines parallèles spécialisées

Utiliser des architectures spécialisées présente deux avantages :

- pouvoir atteindre des performances bien supérieures à celles des architectures générales programmables,
- permettre la réalisation de co-processeurs *peu encombrants et peu onéreux,* accessibles à un grand nombre d'utilisateurs.

En contrepartie, il est clair que le prix à payer se manifeste par un temps de développement plus long et par une évolution limitée.

L'architecture de ces machines, représentée schématiquement sur la figure 1, se compose d'un réseau de N processeurs connectés de voisin à voisin, d'un ordinateur hôte et d'une interface réalisant la liaison entre les deux. L'application principale se déroule sur l'ordinateur hôte. Elle sollicite le réseau de processeurs lorsque des calculs intensifs sont requis. La majorité des machines spécialisées, étudiées ou réalisées à ce jour, possèdent ce type de structure appelé réseau *systolique.*

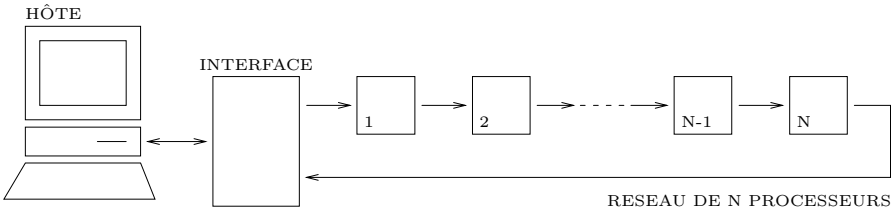

Figure 1. *architecture typique d'une machine spécialisée dans l'analyse de séquences*

Les réseaux systoliques se caractérisent par une topologie régulière, c'est à dire par un maillage uniforme de processeurs connectés localement. Les données sont soumises au réseau à intervalles réguliers, appelés cycles systoliques par analogie aux phases de contraction (systole) et de dilatation (diastole) du coeur et des artères. Ainsi, un cycle systolique consiste en une phase d'émission (de données), d'une phase de calcul et d'une phase de récupération (des résultats). En général, les donnée traversent le réseau en plusieurs cycles systoliques ; à chaque cycle elles progressent d'un pas dans le calcul.

Le calcul des composantes de la matrice se parallélise extrêmement bien sur ce type de réseau, en particulier lorsqu'il s'agit d'explorer une banque complète. La parallélisation s'effectue alors de la manière suivante : la séquence à tester est répartie à raison d'un caractère par processeur et la banque est injectée à une extrémité du réseau. Les éléments de la banque circulent à travers le réseau et rencontrent ainsi tous les caractères de la séquence ; à chaque rencontre un calcul élémentaire est réalisé.

Les performances des machines dépendent à la fois de la complexité des processeurs (ou complexité du calcul à réaliser), du nombre de processeurs et de la fréquence à laquelle la base de données est émise vers le réseau. Les performances maximales (P_{max}) d'une machine en nombre de CME par seconde sont données par la formule :

$$P_{max} = f \times N$$

où f est le nombre de caractères émis par seconde vers le réseau et N le nombre de processeurs.

Les performances réelles d'une machine sont en général moindres. Dans la réalité, il est difficile d'utiliser de manière optimum le réseau de processeurs : la taille d'une séquence test (T) est rarement égale à la taille du réseau (N) ; si elle est plus courte, $(N - T)$ processeurs sont inactifs. Dans ce cas, la quantité de calculs effectués est :

$$f \times (N - T)$$

Si la séquence est plus longue ($T = N \times a + b$), le traitement doit être découpé en $a + 1$ étapes, les a premières étapes utilisant pleinement le réseau et la dernière ne l'utilisant que partiellement.

Le paragraphe suivant décrit cinq machines spécialisées bâties sur ce modèle. Elles ont toutes pour vocation d'accélérer les recherches effectuées sur les banques de séquences biologiques. Cette description n'est pas exhaustive. Elle est cependant représentative de l'état de l'art actuel. Les deux premières font appel à des puces spécialisées conçues spécifiquement pour ces machines ; les deux suivantes intègrent des circuits logiques reconfigurables [ROS 93] dans lesquels un algorithme peut être *câblé* et modifié à volonté ; la dernière est une approche mixte et est un projet de recherche mené conjointement entre l'Irisa à Rennes et l'Inria à Rocquencourt.

4. Quelques machines dédiées à l'analyse de séquences biologiques

4.1. *La machine Bisp*

La machine Bisp [CHO 91] (*Biological Information Signal Processor*), développée à l'Institut de Technologie de Californie, est constituée d'un réseau systolique linéaire spécialisé dans la recherche d'alignements locaux par l'algorithme de Smith et Waterman [SMI 81] présenté au paragraphe 2. Elle constitue un accélérateur matériel connecté à une station de travail.

La machine utilise une puce spécialisée dans laquelle 16 processeurs sont intégrés. Au maximum, 262 144 puces peuvent être cascadées pour former un réseau de 4 194 304 processeurs (!). Un prototype de 256 processeurs (un circuit imprimé de 16 puces) a été réalisé et testé.

La puce permet une recherche de similarité locale ou globale. Plusieurs paramètres comme les coûts associés aux insertions/omissions, le choix de la matrice de substitution, la définition de l'alphabet, ou la précision des fenêtres de recherche sont directement implantés dans le matériel. Une fonction de seuillage est également présente pour filtrer les résultats.

Les performances de la machine dépendent bien sûr du nombre de processeurs. La fréquence d'horloge du circuit étant de 12,5 Mhz et un processeur étant capable d'effectuer un calcul par période, la puissance de calcul est de $12,5 \times 10^6 \times N$ CME/s. Une machine Bisp de 256 processeurs possède une puissance de calcul maximale de $3,2 \times 10^9$ CME/s.

L'interface entre le réseau et la machine hôte est basée sur le microprocesseur MC68020 de Motorola et une mémoire locale de faible capacité utilisée comme tampon d'entrée/sortie. L'alimentation en données s'effectue directement par DMA entre le réseau et la machine hôte. Le taux de transfert est de 3 Mo/s. Le réseau est donc sous-alimenté puisqu'il exige une fréquence d'alimentation de 12,5 M caractères/s.

4.2. *La machine BioScan*

La machine BioScan [SIN 93] est, comme la précédente machine, un système dédié permettant l'accélération d'un algorithme particulier. Il s'agit cependant d'un algorithme moins complexe puisque la recherche d'homologie porte sur des segments de même longueur sans possibilité d'insertion ni d'omission. Cette machine est développée par l'université de Caroline du Nord à Chapel Hill (Etats-Unis).

BioScan est une architecture linéaire de faible complexité. Chaque puce contient 812 processeurs 1 bit et la machine contient 16 puces, soit un ensemble de 12 992 processeurs. Le réseau est connecté via une interface spécialisée au bus Vme d'une station de travail, elle-même accessible par le réseau Internet comme serveur d'applications : les programmes nécessitant une recherche intensive dans les bases de données font directement appel au système BioScan.

Le nombre élevé de processeurs lui confère une grande puissance de calcul. La fréquence d'horloge de la puce est de 32 MHz. A raison de 16 cycles d'horloge pour un calcul matriciel élémentaire, BioScan atteint 25×10^9 Cme/s. Rappelons toutefois que c'est un calcul beaucoup plus simple que dans le cas précédent et que l'unité Cme n'a pas exactement la même signification.

L'alimentation du réseau demande une donnée (un caractère codé sur 5 bits) toutes les 500 ns. La banque de données provient de la machine hôte, les données étant compactées sur 32 bits puis décodées avant d'être émises vers le réseau. La bande passante requise est de l'ordre de 1,25 Mo/s et permet de s'affranchir d'une mémoire locale.

4.3. *La machine Bioccelerator*

La machine Bioccelerator [BIO 93], développée au Weiztmann Institute of Science, en Israël, est une machine spécialisée dans l'accélération de certains programmes du progiciel Gcg [GCG 93] [DEV 84] (Genetics Computer Group). Le coeur de la machine est constitué de circuits logiques reconfigurables [ROS 93] (Fpga : *Field Programmable Gate Array*).

Un Fpga, ou circuit logique reconfigurable, est un composant électronique composé d'une matrice de blocs élémentaires dans lesquels une fonction logique peut être programmée. Ce composant dispose de ressources de routage (également programmables) permettant d'associer ces blocs élémentaires. Ainsi, l'élaboration d'un opérateur arithmétique, un additionneur 16 bits par exemple, consiste à déterminer la fonction logique binaire de l'addition dans 16 blocs élémentaires, puis à spécifier les connexions entre ces blocs. Une architecture matérielle est constituée d'un assemblage d'éléments spécifiés de la sorte. La programmation d'un tel composant est entièrement dynamique et s'effectue en quelques millisecondes.

La machine Bioccelerator se compose d'un système de mémorisation de grande capacité permettant de stocker en interne les banques de séquences à

traiter et de une à quatre cartes contenant chacune quatre noeuds de calcul. Un noeud contient un FPGA Xilinx 4008 et une mémoire statique rapide et permet la mise en oeuvre d'un processeur 24 bits.

La machine n'est donc pas dédiée à un algorithme particulier : en modifiant les configurations associés aux FPGAs, les structures de la machine peuvent être directement adaptées à un algorithme donné. Par contre, le temps d'implantation et de mise au point d'un nouvel algorithme requiert les compétences d'un architecte de machines.

Bioccelerator est commercialisée ; dans sa première version, elle accélère le programme *ProfileSearch* du progiciel GCG. Les performances mesurées sur ce programme particulier sont de 320×10^6 CME/s dans sa configuration maximale.

Grâce à la grande capacité de la mémoire interne les processeurs sont constamment alimentés et travaillent en permanence. Ce très bon équilibre entre puissance de calcul et capacité d'alimentation fait que les performances réelles sont très proches des performances théoriques.

4.4. *La machine Splash-2*

La machine Splash-2 [ARN 92], conçue au SRC (*Supercomputing Research Center - Institut for Defense Analyses*), présente d'intéressantes possibilités. C'est également un système à base de circuits programmables (FPGA). Elle est le successeur de la machine Splash-1 [LOP 91] initialement étudiée pour des applications en biologie moléculaire.

Splash-2 est un réseau linéaire dont chaque noeud est composé d'un FPGA (Xilinx 4010, l'équivalent de 10 000 portes logiques) et d'une mémoire statique de 512 Ko. La configuration maximum est de 256 noeuds répartis sur 16 cartes distinctes. L'interfaçage avec une station de travail (Sun Sparc Station) est assurée par le SBus (bus spécialisé dans les entrées/sorties) et autorise un débit maximum de 54 Mo/s.

L'intérêt de Splash-2, tout comme la machine Bioccelerator est de pouvoir *programmer* une architecture matérielle. Cette machine possède les performances des architectures dédiées, mais peut être programmée. La programmation est cependant difficile et nécessite l'intervention d'un expert en architecture de machine pour l'implémentation d'algorithmes.

Les premières mises en oeuvre d'algorithmes d'analyse de séquences [HOA 93] montrent qu'il est possible d'implanter plus d'un processeur par noeud. On peut estimer à deux le nombre de processeurs supportant l'algorithme de Smith et Waterman et pouvant être contenus dans un noeud, soit un réseau de 512 processeurs. Avec une fréquence d'horloge de 20 MHz, la machine atteint 10×10^9 CME/s. La bande passante de 54 Mo/s ne constitue pas un facteur limitatif.

4.5. *La machine Samba*

La machine Samba (*Systolic Accelerator for Molecular Biology Application*) est un projet d'architecture pour l'analyse des banques de séquences développé à l'Irisa. Samba est une architecture parallèle linéaire. Elle est constituée de puces spécialisées pour la partie réseau et de FPGAs pour la partie contrôle et alimentation en données. L'algorithme de base supporté par les processeurs est l'algorithme de Smith et Waterman.

La puce est actuellement en cours de conception. A terme. elle devrait contenir plusieurs processeurs 16 bits et fonctionner à une fréquence d'horloge de 20 MHz. Une machine prototype de 256 processeurs est prévue. Les performances du réseau se situent donc aux alentours de 5.12×10^9 CME/s ($256 \times 20 \times 10^6$).

Le lien avec une station de travail est réalisée (dans la version prototype) par une carte expérimentale à base de circuits logiques reconfigurables (FPGA), la carte Perle1 développée à DEC-PRL [BER 92] [BER 93]. Cette carte comporte une matrice de 4×4 FPGAs (Xilinx 3090), une mémoire statique rapide de 4 Mo répartie tout autour de la matrice, une interface TURBOchannel offrant une bande passante de 100 Mo/s et des possibilités d'entrées/sorties vers le monde extérieur de l'ordre de 800 Mo/s.

Ces ressources permettent de *câbler* une interface performante assurant simultanément l'alimentation en données et la récupération des résultats. Ces derniers, une fois filtrés, peuvent être soit stockés localement, soit transmis à l'hôte. Le débit élevé procuré par le TurboChannel entre l'hôte et la carte Perle-1 devrait permettre de s'affranchir d'une mémoire locale de grande capacité.

Cette approche mixte (puces spécialisées/FPGAs) est intéressante. Elle permet d'abord d'exploiter la régularité de la structure linéaire en intégrant plusieurs processeurs par puce. On obtient ainsi un réseau de grande taille dans un minimum de volume. D'autre part, la technologie des circuits logiques reconfigurables permet d'optimiser l'interfaçage entre l'ordinateur hôte et le réseau sans le figer complètement. Suivant les problèmes à résoudre (comparaison de deux séquences, d'une séquence avec une banque, de deux banques), le contrôle du réseau exige des mises en oeuvre différentes qui, pour être efficaces, doivent être matériellement adaptées.

5. Discussion

Ce paragraphe tente d'établir quelques comparaisons entre les cinq machines présentées précédemment. Le premier élément de comparaison qui peut être pris en considération concerne les performances maximales du réseau. Rappelons que ces performances sont déterminées par le produit du nombre de processeurs et de la fréquence d'horloge du réseau.

Cette mesure *brute* peut ensuite être raffinée en évaluant les performances du système complet. On ne considère plus la fréquence d'horloge du réseau mais la fréquence avec laquelle les données peuvent parvenir au réseau.

Dans les deux cas, l'unité de mesure est le calcul matriciel élémentaire (CME). Le tableau ci-dessous rappelle ces valeurs estimées (en M CME/s) au paragraphe précédent. Tous les processeurs sont considérés comme actifs.

	Bisp	BioScan	Bioccele.	Splash-2	Samba
réseau	3200	25000	320	10000	5120
système	800	25000	320	10000	5120

Rappelons qu'un CME de la machine BioScan est de plus faible complexité qu'un CME des autres machines. En règle générale, les machines dédiées sont en mesure de tenir leur performance grâce à une alimentation en données adéquate. Le réseau de la machine Bisp est cependant sous exploité à cause de la faible bande passante avec l'hôte et de la capacité réduite de la mémoire sur l'interface.

Une autre manière de comparer ces machines est d'estimer leur temps de réponse lorsqu'elles sont soumises à une requête particulière. Supposons que l'on veuille confronter une séquence de taille T avec une banque équivalente de 10×10^6 caractères (banque protéique de 30 000 séquences environ). Le temps de réponse est calculé approximativement de la manière suivante [1] :

$$t = \frac{10^7}{f} \times \left(\left\lceil \frac{T}{N} \right\rceil + 1 \right)$$

où f est la fréquence d'émission des données vers le réseau et N la taille du réseau. Le diagramme de la figure 2 illustre les temps d'exécution des cinq machines ; ils progressent par palier et sont directement fonction de la taille de la séquence test. La machine Bioccelerator, de part le nombre réduit de processeurs qu'elle contient, présente un handicap dès que la taille des séquences à traiter augmente. Elle reste quand même intéressante par rapport à une solution programmable (de type MasPar, par exemple) car elle offre sensiblement les mêmes performances pour un coût très inférieur. De plus, c'est la seule machine commercialisée actuellement avec un jeu de programmes directement exploitable dans l'environnement GCG.

A l'opposé, la machine Splash-2 détient le record. Celui-ci se paye par un *volume de matériel beaucoup plus important* (un facteur 10 à 20 environ). Un réseau de 256 processeurs (équivalent Bisp ou Samba) implémenté sur la machine Splash-2 nécessite plusieurs cartes complexes représentant l'équivalent de plusieurs centaines des meilleurs composants actuels (Xilinx 4010, mémoires statiques rapides, routeurs spécialisés, ...). Le coût d'une telle machine est du même ordre de grandeur que celui d'une machine parallèle programmable. Splash-2 est commercialisée mais ne propose aucun logiciel relatif au domaine de la biologie moléculaire.

Les machines Bisp, BioScan et Samba se situent entre ces deux extrêmes. Elles ont en commun un réseau de processeurs VLSI spécialisés tenant sur un

[1] $\lceil x \rceil$ désigne la partie entière supérieure de x

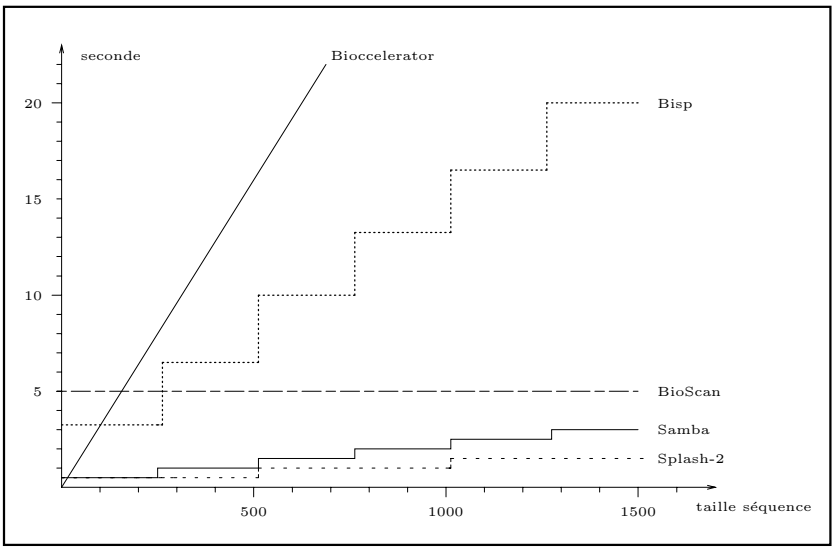

Figure 2. *temps d'exécution d'une comparaison sur une base de données de* 10^7 *caractères en fonction de la taille de la séquence test*

seul circuit imprimé. Leur coût et leur encombrement sont faibles. Actuellement, aucune de ces machines n'est commercialisée.

6. Conclusion

Les algorithmes couramment utilisés pour l'analyse des banques de séquences se prêtent particulièrement bien à une parallélisation sur des réseaux linéaires de processeurs. Ces algorithmes étant coûteux, en terme de calcul, des machines parallèles spécialisées sont indispensables pour faire face à l'accroissement extrêmement rapide des banques de séquences.

Plusieurs machines, principalement bâties autour de structures linéaires, ont d'ores et déjà été étudiées et développées. Certaines sont même opérationnelles comme la machine BioScan (connectée à un serveur accessible par réseau) ou la machine Bioccelerator (commecialisée par la société Compugen, Israël).

Ces machines sont conçues autour de puces spécialisées et/ou de circuits logiques reconfigurables (FPGAs). L'intégration de plusieurs processeurs par puce permet d'obtenir des machines de faible dimension (un circuit imprimé). Connectées à une station de travail standard, elles sont nettement plus performantes que les machines parallèles programmables.

Ces machines visent principalement deux secteurs d'utilisation. Le premier concerne la mise à disposition d'une ressource performante à une communauté de personnes ayant des besoins ponctuels. On peut imaginer une connexion à un

serveur offrant immédiatement des services impossibles à réaliser localement. Dans ce cas, le centre serveur dispose d'une ou plusieurs machines spécialisées qu'il active en fonction des requêtes.

L'autre secteur d'utilisation concerne l'usage intensif de telles machines. Les recherches menées, par exemple, sur la classification d'une banque de séquences par analyse d'homologies [SON 94] demandent une masse de calculs gigantesque. Dans ce cas, l'exploitation permanente (ou sur un temps très long) d'une ressource d'un centre serveur n'est pas envisageable. Le faible coût d'une machine spécialisée (comparativement à une machine parallèle programmable qui offrirait les mêmes performances) permet à une équipe de recherche de se doter d'une telle machine, voire de plusieurs.

7. Bibliographie

[COD 91] Codani JJ. and Lacroix B., "Computational aspect of genome physical mapping", research report 1560, INRIA, 1991.

[BEL 92] Bellanné-Chantelot et al., "Mapping the whole human genome by fingerprint yeast artificial chromosomes", *Cell*, vol. 70, p. 1059-1068, 1992.

[NIC 90] Nickolls J.R., "The Design of the MasPar MP-1: A Cost Effective Massively Parallel Computer", *COMPCON*, p. 25-28, 1990.

[LIP 85] Lipman R.J. and Pearson W.R., "Rapid and sensitive protein similarity searches", *Science*, vol. 227, p. 1435-1441, 1985.

[PEA 88] Pearson W.R. and Lipman R.J., "Improved tools for biological sequence comparison", *Proc. Natl. Acad. Sci.*, vol. 85, p. 3244-3248, 1988.

[ALT 90] Altschul S.F et al., "Basic Local Alignment Search Tool", *J. Mol. Biol.*, vol. 215, p. 403-410, 1990.

[SON 94] Sonnhammer E.L. and Kahn D., "The Modular Arrangement of Proteins as Inferred from Analysis of Homology", *Protein Science*, to appear, 1994.

[LEM 93] Lemoine E., "Reconfigurable Hardware for Molecular Biology Computing Systems" *ASAP'93*, IEEE Computer Society Press, p. 184-187, 1993.

[SMI 81] Smith T.F. and Waterman M.S., "Identification of common molecular subsequences" *J. Mol. Biol*, vol. 147, p. 195-197, 1981.

[ROS 93] Rose J. et al., "Architecture of Field-Programmable Gate Arrays", *Proceedings of the IEEE*, vol. 81, n^o 7, p. 1013-1029, Jul 1993.

[CHO 91] Chow E. et al., "Biological Information Signal Processor" *ASAP'91*, p.144-160, 1991.

[SIN 93] Singh R.K. et al., "A Scalable Systolic Multiprocessor System for Analysis of Biological Sequences", *Research on Integrated Systems*, p. 168-182, 1993.

[BIO 93] "The BIOCCELERATOR machine", *Documentation Technique*, 1993.

[GCG 93] Genetics Computer Group, "Program Manual for the GCG package, version 7", 1993.

[DEV 84] Devereux J. et al., "A comprehensive set of a sequennce analysis programs for the Vax", *Nucl. Acids Res.*, vol. 12, p. 387-395, 1984.

[ARN 92] Arnold J.M. et al., "SPLASH 2", *4th Annual ACM Symposium on Parallel Algorithms and Architecture*, 1992.

[LOP 91] Lopresti D., "Rapid Implementation of a Genetic Sequence Comparator Using FPGAs", *Advance Research in VLSI 1991*, p. 139-152, 1991.

[HOA 93] Hoang D.T., "Searching Genetic DataBases on SPLASH-2", *FPGAs for custom computing machines*, IEEE Computer Society Press, p. 185-191, 1993.

[BER 92] Bertin P. et al., "Programmable Active Memories : a performance assessment", *Parallel Architectures and their efficient use*, Lecture notes in Computer Science, Springer-Verlag, p. 119-130, 1992.

[BER 93] Bertin P., "Mémoires actives programmables : conception, réalisation et programmation", thèse de doctorat, Université Paris 7, 1993.