計画研究：2000～2004年度

ゲノムデータベースからの知識発見

●森下　真一[1)]　◆後藤 修[2)]　◆土井 晃一郎[3)]　◆瀬々潤[3)]

1) 東京大学　大学院新領域創成科学研究科　情報生命科学専攻　　2) 京都大学 大学院情報学研究科　知能情報学専攻
3) 東京大学　理学部　生物情報科学 学部教育特別プログラム

〈研究の目的と進め方〉

ゲノム配列の決定により，遺伝子配列に関する知識はより完備な状態に近づくと考えられる．われわれは，平成12年度より一貫して，大規模なゲノムデータを高速に解析し，知識発見を促すようなソフトウエアについて研究している．主なテーマ，研究期間，研究者（助手，大学院生，分担者）をしめす：

- ●高感度ゲノミックＰＣＲプライマ設計技術（平成14-16年度: 山田智之，森下真一）
- ●ヒト／マウス遺伝子阻害のための siRNA配列設計ソフトウエア （平成15-16年度: 山田智之，森下真一）
- ●5’SAGE解析ソフトウエア （平成15-16年度: 笠井康弘，森下真一）
- ●ゲノム配列を解読するWhole Genome Shotgun Assembler の研究開発（平成14-16年度: 笠原雅弘，佐々木伸）
- ●メダカゲノム概要配列決定 （平成14-16年度: 笠原雅弘，佐々木伸）
- ●5’SAGE法を使った遺伝子推定（平成16年度: バドルル・アーサン，森下真一）
- ●比較ゲノムのための高速アラインメントソフトウエア （平成12-16年度: 山田智之，小笠原準，森下真一）
- ●比較ゲノムとゲノム進化 (平成16年度:中谷洋一郎，曲薇，森下真一）
- ●新規繰返し配列の推定（平成16年度: 山田智之）
- ●ゲノム配列をアノテーションするためのゲノムブラウザー　（平成14-16年度: 橋本順之，永安佑希允，土井晃一郎）
- ●遺伝子発現量等のデータから機能予測するデータマイニング技術（平成13-16年度: 瀬々潤，森下真一）
- ●蛋白質コード領域予測技術（平成13-16年度: 後藤修）
- ●ショウジョウバエの翅脈解析ソフトウエア （平成16年度: 中村恵理，初田浩志，森下真一）

である．ゲノム 4 領域の研究者との共同研究は可能な限り推進してきた．

〈研究開始時の研究計画〉

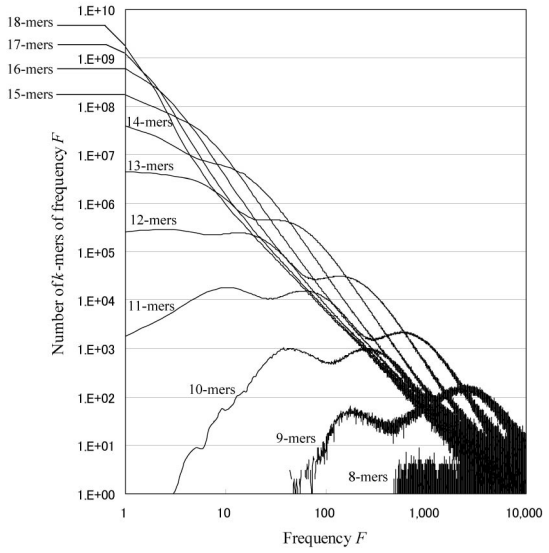
平成12年度に計画研究を開始した段階では，遺伝子発現量データを用いて遺伝子機能を予測する技術を構築することに焦点をおいており，ゲノム解読および比較ゲノムについては漠然とした研究計画しか描けない段階であった．しかしその後ゲノム配列が次々と解読され，whole genome shotgun 方式も注目されるようになり，年度が進むにしたがって解決しなければならない問題が次々と具体化した．このため特定領域研究「ゲノム」に参加している様々な班員との共同研究が続々と開始され，結果として研究目的に挙げたテーマに取り組むこととなった．

〈研究期間の成果〉

- ●　高感度ゲノミックＰＣＲプライマ設計技術
  （平成14-16年度）

ヒトゲノム配列がほぼ決定されたのを受け，長さが20から100程度の部分配列が全ゲノム中で，どの程度ユニークであるかを，

頻度分布により見積もることを試みた．下図にヒトゲノムにおける長さ8から18塩基の短い配列の分布図を示す．



ヒトゲノムにおける長さ8から18塩基の短い配列の分布図：横軸が出現頻度，縦軸がその頻度を持つ配列の総数を表示

その部分配列が全ゲノム配列で唯一存在し，かつk個程度変異を入れても依然としてゲノム配列中に出現しないというミスマッチ耐性(具体的には上記の k)という概念を提唱した．しかしながらミスマッチ耐性を正確に計算することは非常に時間がかかる困難な問題である．そこでミスマッチ耐性は少なくともこれ以上という下限値を計算する近似的解法に取り組んだ．さまざまなシード配列を使って真の値に近い近似解を計算できる限界について解析し，実用的な時間で動作する動的計画法にもとづくアルゴリズムを考案した6)9)．またその成果を平成14年度から web server GenomeSurf として公開している26)．この成果は，ゲノム上にマーカを設計し，オリゴDNA チップの設計に利用されている．また，このソフトウエアを用いて，藤山研究室との共同研究でヒトY染色体上での稠密なゲノムマーカーチップを設計した．

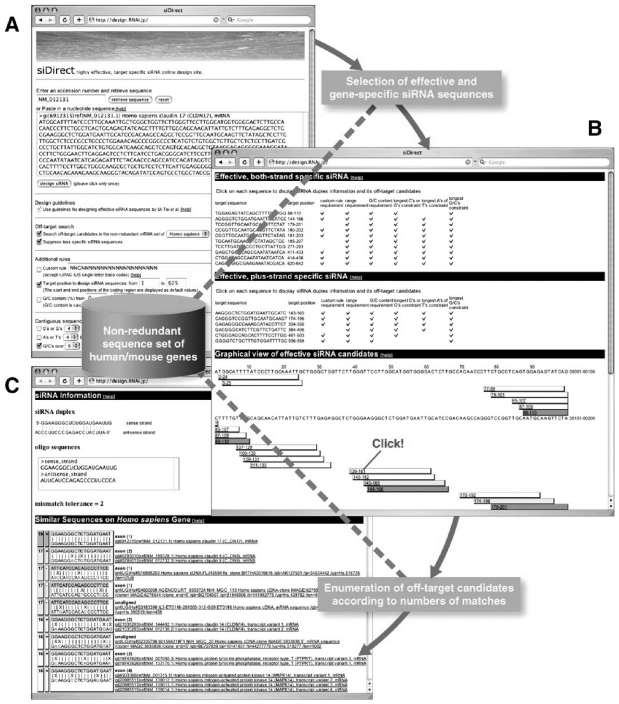
GenomeSURF: ゲノム上の短い配列の頻度分布を使ったゲノムマーカ作成システム http://surf.gi.k.u-tokyo.ac.jp 青く色をつけた部分が特異的な配列を，赤い部分が反復配列を表示している

- ●ヒト／マウス遺伝子阻害のための siRNA配列設計ソフトウエア （平成15-16年度）

siRNA はmRNAの阻害法として急速に広がっいる．標

的とする遺伝子を阻害する効果の高いsiRNA配列設計法は西郷・程研究室が世界をリードしている．一方，siRNAの長さは約21塩基と短いため，標的遺伝子とはまったく無関係な遺伝子にも相同部分が存在する確率が高い．そのような場合，siRNAによって標的とは無関係な遺伝子の発現まで抑制される可能性がある．この現象はsiRNAのオフターゲット効果と呼ばれており，siRNAを用いた遺伝子機能解析や，siRNAを医薬に応用するうえで深刻な問題のひとつとなっている．オフターゲット効果を回避するためには，標的以外のすべての遺伝子に対して相同性が最小となるsiRNA配列を選択することが望ましい．一般的には，設計したsiRNA配列をBLASTで検索し，無関係な遺伝子がヒットしないものを選択すればよいとされている．しかし，ひとつひとつのsiRNA配列をBLAST検索するのは時間と労力がかかるうえ，siRNAのような短い配列の検索では見落としが多いという重大な欠点がある．このためBLASTでsiRNAの特異性を確認する従来の方法では，オフターゲット効果のリスクを過小に評価してしまうことになる．

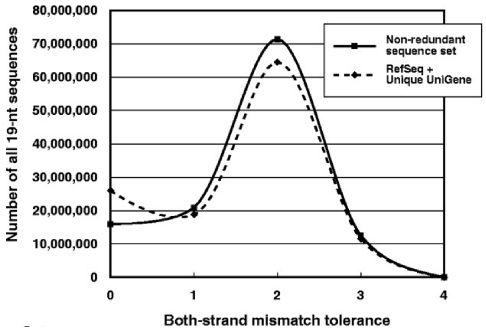
そこで我々は，siRNAのような短い配列の相同性検索を高速かつ確実に実行できるプログラムを開発した1)．標的以外の遺伝子を阻害しないこと（クロスリアクトの防止）を保障するためには，標的以外にはマッチする可能性が少ない配列の設計が重要である．しかし，これを保障するには，全遺伝子配列の走査を伴うため計算時間がボトルネックとなる．この目的で開発されてきた従来法（BYP法，Lee-Sung 法）ではオンラインで計算するのには十分な速度が出ないと言う問題点があった．そこで我々は従来法より一桁高速なアルゴリズムを開発し，実用的な時間でクロスリアクトを防止する配列決定を可能にした1)．設計サーバーは　http://design.rnai.jp/　から公開中である5)21)22)．



siDirect によるsiRNA配列設計の流れ

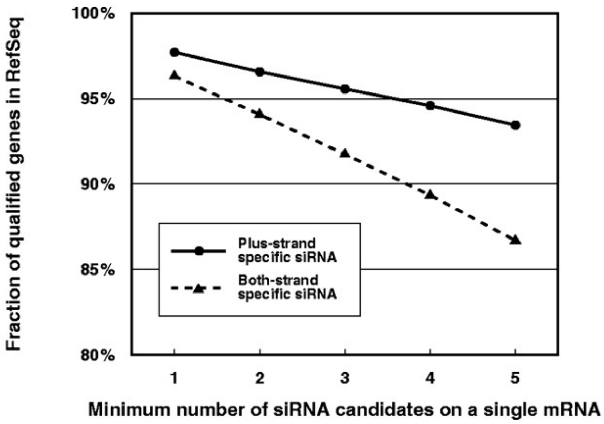
siDirectのトップページを上図Aに示す．この画面上に標的遺伝子の塩基配列またはAccession番号を入力するだけで，効くsiRNA設計のガイドラインを満たす配列のリストが得られる(上図B)．さらに，上部のリストおよび青色のバーで表示されたsiRNAは，センス鎖・アンチセンス鎖とも，本来の標的以外のあらゆる配列に対して必ず3塩基以上のミスマッチが存在する配列である．さらにsiDirect では設計した siRNA 配列とクロスリアクトする危険性のあるミスマッチ数 3 以下の配列とミスマッチする箇所ををすべて表示して注意を喚起している（図C)．

ヒトのRefSeq mRNAデータベースにおいては，必ず3塩基以上のミスマッチが存在する 19 塩基の配列は全体の約10%存在するが，4塩基以上のミスマッチを保証できるものはほとんど存在しない（下図)．したがって，siRNA設計においては3塩基以上のミスマッチがある配列を選択するのが最善といえる



ミスマッチ耐性の分布図　横軸がミスマッチ耐性，縦軸がそのミスマッチ耐性をもつヒト遺伝子上の 19塩基配列の総数5)

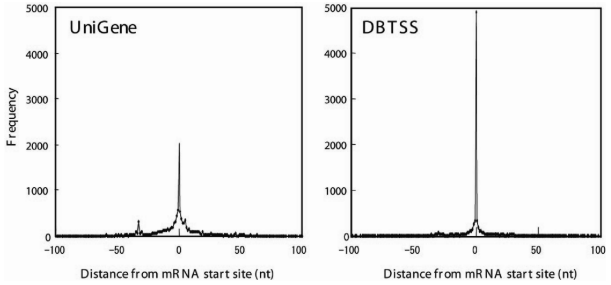
するとsiRNA配列設計問題は，オフターゲットに対しては必ず3塩基以上のミスマッチが存在する19塩基配列から，siRNAの基準を満たすものを各遺伝子について設計できるか否かが鍵となる．この可能性をヒト全遺伝子に対して調査した結果が下図である．基準を満たす siRNA配列が少なくとも一つ存在したヒト遺伝子は全体の96%を占めることがわかり，本手法のように厳しい基準で設計してもオフターゲット効果が少ない配列を設計できることが理解された．



少なくとも1個以上のsiRNA 配列が設計できるヒト遺伝子の割合：　横軸は k 個以上の siRNA 配列が設計できること，縦軸はそのようなヒト遺伝子の割合を表示5)

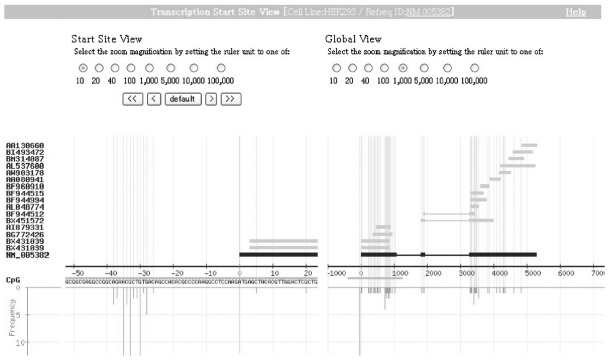
● 5’SAGE解析ソフトウエア （平成15-16年度）

ヒト遺伝子を対象とした5’SAGE法は転写開始地点の頻度分布をハイスループットで観測できる方法として注目されてきたが，平成15年暮，橋本真一博士が研究開発に成功した4)．



5’SAGE により観測された転写開始点と従来の UniGene および DBTSS により同定されていた転写開始点の間の距離：従来の転写開始点に比べ -500 から +200 の領域に入った 5’SAGEタグの割合は UniGene の場合，98.2% DBTSS では 96.4% であり，5’SAGEタグによる転写開始点の推定が高い確率で正しいことが理解された．

5’SAGE情報はハイスループットでヒトゲノム中での頻度分布の解析を可能にする．解析の効率化を目指してソフトウエア 5’SAGEを研究開発し公開した2)．



5’SAGE による転写開始点の表示： オレンジ色の線が各5’SAGEタグの出現位置をx座標で表示し，下に延びた線の長さが出現頻度を表現

● ゲノム配列を解読するための Whole Genome Shotgun Assembler の研究開発（平成14-16年度）

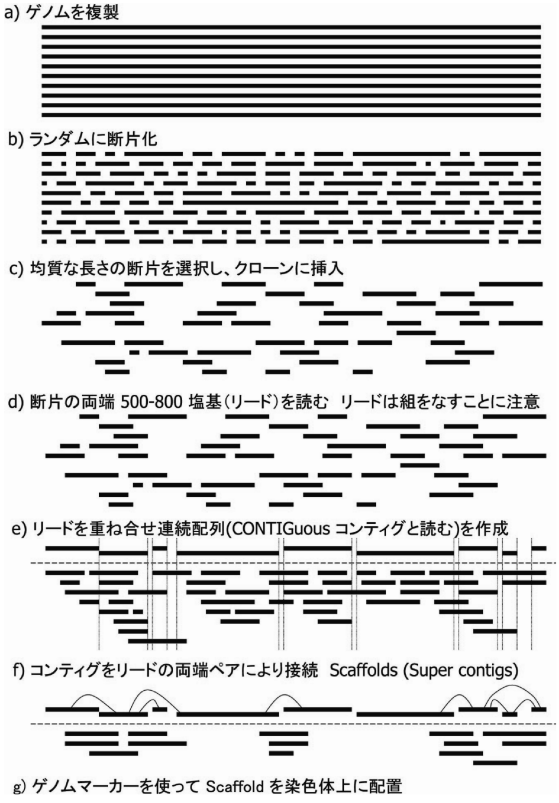
1982年Sangerがショットガン・シークエンシング法を用いてbacteriophage lambda のゲノム配列を決定して以来，より長いゲノム配列がこの方法を改良することにより解読されてきている．現在，ヒトゲノムの解読を契機に，さまざまな脊椎動物のゲノム解読が進行している．対象とするゲノムが大きくなるにしたがって，人手で解読する比重は低くなり，解読ソフトウエアであるゲノムアセンブラの役割が重くなる．

短い配列しか読めない技術を使って長大なゲノム配列を解読するには，どのような工夫が必要か？ 自然な考え方は，ゲノムを短いDNA断片へと分解し，各DNA断片を解読するという方法が考えられる．このとき各DNA断片の間に重なりがないと，DNA断片がゲノム上でどのような順番で繋がっているかどうかがわからない．そこでDNA断片同士の順序を決めるために，断片同士が重なるように冗長な断片を作成し解読すればDNA断片を繋げ合わせて伸ばしてゆける．この方法論をショットガン・シークエンシング (shotgun sequencing) と呼ぶ．

サンガーによるbacteriophage lambdaゲノム解読の成功は，ショットガン・シークエンシングを利用したゲノム配列解読への関心を高めた．しかし，10万塩基対程度のウイルスゲノムに比べて遥かに長いゲノム配列の決定に有効かどうかは長らく疑問視されていた．注意しなければならない問題がいくつかある．真核生物のゲノムには，レトロポゾンが生み出した数多くの繰り返し配列が存在する．たとえば，ヒトゲノムには長さ約280塩基対のAluと呼ばれる繰り返し配列が約100万コピー存在し，1000塩基対を超える繰り返し配列も多種類存在する．DNA断片が共有する配列が繰り返し配列により覆われ，繋ぎ合わせ方を一通りに決めることができない場合が生じる．このような曖昧な状況では，アセンブリミスを回避するため，繋ぎ合わせを止めなければならない．

繰返し配列の影響を避けるための典型的方法は，ゲノム全体を一度に断片化するのではなく，ある程度の長さの区画に分割して各区画を個別に解読するアプローチであり，階層型アセンブリ(hierarchical assembly)もしくは区画化アセンブリと呼ばれる．具体的には各区画を長さが10万から20万塩基対程度の区画に分割し，バクテリア人工染色体(BAC)の中に入れてクローン化する．たとえばヒトゲノム中に存在する約100万コピー存在する繰返し配列Aluも，約2万個のBACで区画すれば，1 BACあたり平均約50コピーにまで減少する．このように各区画内での繰り返し配列が生む曖昧性をかなり排除しアセンブルしたのちに，区画をつなぎ合わせるのが階層型アセンブリである．しかしながら区画化はコストが高く，多くの人的労力を必要とする．

ゲノムを区画化するためのコストと人的労力は大きい．区画化せず，ゲノム全体を一度に断片化し，断片配列を解読し，ソフトウエアによりつなげ，元のゲノムの大まかな構造を再構成するホールゲノム・ショットガン・シークエンシング(whole genome shotgun sequencing；WGS)を実現できれば，大幅なコスト削減を期待できる．



ホールゲノム・ショットガン・シークエンシング方式の原理

1995年 TIGR(The Institue of Genome Research) はホールゲノム・ショットガン・シークエンシングにより Haemophilus influenzae ゲノムの1,830,137 塩基対を解読した．しかし，解読を完成させるまでには多くの人手による追加実験が必要であった．また，数千万塩基対を超える真核生物のゲノム配列には多くの繰返し配列が存在するため，ホールゲノム・ショットガン・シークエンシングを使用することは本質的に困難であると Green は主張した．一方 Venter らは，ランダムに生成したゲノム断片をプラスミド，コスミド，BACベクターに挿入し，挿入配列の両端配列の組を情報として取り入れることにより，ヒトゲノムを解読できる可能性を示唆した．長い論争の末，2000年Celera社はホールゲノム・ショットガン・シークエンシングにより，Drosophila melanogaster のゲノム配列（真正染色質の約1.2億塩基対）を解読し，つづいて2001年にヒトゲノムの解読にも成果を収めた．この結果ホールゲノム・ショットガン・シークエンシングは，階層型アセンブリに比べ誤りは多いものの，低コストでゲノムの大まかな構造を決めるのに有用であると認識された．

このようにして階層型ゲノムアセンブリに比べ安価なコストでゲノムを解読できる whole genome shotgun assembly が広まった．ヒトやマウスのゲノム解読に利用されている．今後も様々なゲノムがこの方式で解読されてゆくと考えられるが，高精度のアセンブリを実現するためには解決しなければならない問題も多いことが研究調査の結果わかった．

まず脊椎動物や高等植物ゲノムでは繰り返し領域が多数存在し，解読を撹乱するのでmate-pair 情報をうまく利用することが成功のポイントになる．また純系でないサンプルではしばしばハプロタイプ間の相同性が低い領域が存在し，リードのカバー率が低い場合には解読が困難である．この問題を低コストで解決するには，リード数をあまり増やさずにカバー率を最大化したい．そこで，ベクター配列やコンタミなど不必要な部分だけを丁寧に除去したいが計算コストが高くつく．典型的な脊椎動物のゲノム解析では，数百万から数千万個のリードを処理する必要があるが，計算性能をあげる工夫をしないと，この処理だけで数十日の計算時間を消費することになる．

また，アセンブル結果が正しいか否かを精査し，不具合を編集する機能をそなえた専用エディタも必要である．しかし数百万個のリードからなるスーパーコンティグを高速に表示し分析できるようにするには，適切なデータ圧縮技術で計算資源を上手に使った設計をしなければならない．

このような問題点を解決することを目指してwhole genome shotgun assembly 方式の研究に取り組んだ．従来型のソフトウエアである Phrap, Arachne 等のアルゴリズムのマイナーな改良する方針はとらずに，個々のステップを全てオリジナルな考え方で設計したソフトウエア Ramen assembler を研究開発した．

まず我々が注目したのはメイトペア情報を利用したレイアウトとコンティグ伸張である．DNA断片をクローン化する際，ベクターへ挿入可能なDNA断片の長さは，ベクターによって異なる．たとえば，プラスミド，コスミド，BACのベクターには，各々 1万塩基対ぐらいまで，3.5〜4.5万塩基対，10〜20万塩基対程度のDNA断片を挿入しクローン化できる．ベクターに挿入されたDNA断片の両端を端から長さ1000塩基対程度の配列は読むことができ，２つの両端配列を組として認識できる．この組のことをメイトペア(mate-pair)もしくはエンドペア(end-pair)と呼ぶ．たとえば4000塩基対前後のDNA断片をクローン化できるプラスミドベクターの両端配列を各々1000塩基対読めば，両端配列に挟まれ解読できなかった中間部分の長さは2000塩基対前後と推定できる．かりに未解読の中間部分に繰り返し配列が含まれていたとしても，それを飛び越え，両端配列と重なるDNA断片を使って両端のコンティグ配列を伸ばせる可能性が高い．このようにメイトペア情報は繰返し配列によるアセンブリの誤りを回避するのに役立つ．

つづいて改善したのはスッキャフォールド生成である．メイトペア情報を利用して伸張させて得られる配列は必ずしも連続したDNA配列でなく，中間には解読できなかった部分がギャップとして残る場合が多い．このような状態をコンティグとは別の言葉スキャッフォールド(scaffold) で表現する．スキャッフォールド生成にはメイトペア情報が鍵となるが，キメラの存在を考えると慎重に使わなければならない．たとえば１つのメイトペア情報だけで，２つのコンティグ・スキャッフォールドを接続するとキメラによる誤った接続が生成される可能性もある．そこで複数の独立のメイトペアにより繋がるかどうかを調べる．1本のメイトペアしかない場合には，繋ごうとしている配列が近くにあることを示す他の強い証拠（遺伝距離など）を鑑みて慎重につなぐ．

現在の技術ではメイトペアは短いものほど収集しやすい．そのため，短いメイトペアを使ってスキャッフォールドを伸張させながら，近くに存在するメイトペアと矛盾した位置を繋げようとするメイトペアを見つける．このようなメイトペアはキメラの可能性があるため除くことが無難である．一方，メイトペアに挟まれた未読の領域は，その長さを概算することが可能になるため，囲む周辺配列を伸ばすことで未読領域を推定することが容易になる．スキャッフォールド長が1万塩基対前後になったところで，長さ1万塩基対のメイトペアを使うと，キメラの排除をしながら，スキャッフォールドを効果的に伸ばすことができる.4万塩基対程度の長さになったところで，3.5〜4.5万塩基対をクローン化可能なフォスミドベクターのメイトペアを使ってスキャッフォールド同士を繋ぐことで，より長くする．さらに20万塩基対程度の長さになった段階でBACのメイトペアを利用する．段階的にメイトペアを利用することで100万塩基対を超えるスキャッフォールドへ伸ばすことも可能である．

● メダカゲノム概要配列決定 （平成14-16年度）

小原研究室がシークエンシングした whole genome shotgun read を繋げ，メダカゲノムの概要配列を決定した．アセンブル可能なゲノムサイズを700.4M塩基と推定し，この母数に対して 10.55倍の重複度で解読したreadを繋げた．ゲノム全体の50%を覆うのに必要な最小の長さ，いわゆる N50値は scaffold で1.41 M 塩基，ultra-contig で 5.1M 塩基となり十分な長さでゲノムを被覆することができた．また染色体中での scaffold の位置を推定するために，ゲノムマーカを武田研究室とともに設計し，約90% をゲノム上に配置することに成功した．既に解読された10個のBAC 配列との比較を行なった結果，塩基レベルでのエラー率は 0.17%（リードの両端100塩基は低QV値のため除くと 0.05%）という極めて良好な精度を得ている．また平成１５年暮れからカイコゲノムの概要配列をアセンブリした7).

type	size	sequence coverage	clone coverage
Plasmid	2.6kb	9.69	23.1
Plasmid	7.5kb	0.48	4.3
Fosmid	35.5,37.5kb	0.26	11.1
BAC	130-210kb	0.12	15.0
Total		10.55	53.5

メダカゲノムアセンブリに用いたクローンライブラリの構成と塩基被覆率とクローンによる被覆率

		bases (Mb)	Percentage
Anchored	oriented	584.0	83.37
	unoriented	14.7	2.10
	unordered	29.1	4.16
Unanchored		72.6	10.37
Total		700.4	100.00

ゲノム上に配置することができた総塩基数：方向（ストランド）が決まったもの，決まらなかったもの，おおよその位置が推定できたものの scaffold の順番が未定なものに分類

●5’SAGE法を使った遺伝子推定（平成16年度: バドルル・アーサン，森下真一）

ゲノムから遺伝子をコードしている領域を自動的に推定する遺伝子予測の研究の歴史は長い．コドンの頻度分布，イントロンとエキソンの境界領域に保存されているモチーフ配列，ＧＣ率を考慮し隠れマルコフモデルを用いた遺伝子予測手法は一般的であり，エキソン領域の予測にはある程度の成果をあげている．しかしながら転写開始点を予測し，5’UTR 領域を予測することは頼りになるモチーフが少ないため，難しい問題として残されてきた．この問題を解決する直接的な手段は，転写開始点だけを実験により効率的に収集してしまうことである．

我々はメダカの初期胚および成体から1,186,742個（重複を除くと392,341個）の5’SAGEタグを解析し，そのうち1,069,807個（重複を除くと320,474個）がゲノム上にアラインメントされた．このようにして得られた転写開始点から GeneScan による遺伝子予測を行うことを試みたが，そのままででは予測が困難であることが分かった．最も難しい問題は GeneScan が第一エキソンと翻訳開始点を出力しない場合があるとことであった．そこでこの問題を補うため，第一エキソンと翻訳開始点を推定するアルゴリズムを開発した．そのメダカゲノム解析からは以下に示す結果が得られた．

	Number
予測遺伝子数	20,141
総エキソン数	158,219
エキソン数 / 遺伝子	7.8
コード領域の総塩基数	28,485,099bp
エキソンの平均長	180bp

●比較ゲノムのための高速アラインメントソフトウエア (平成12-16年度: 山田智之，小笠原準，森下真一)

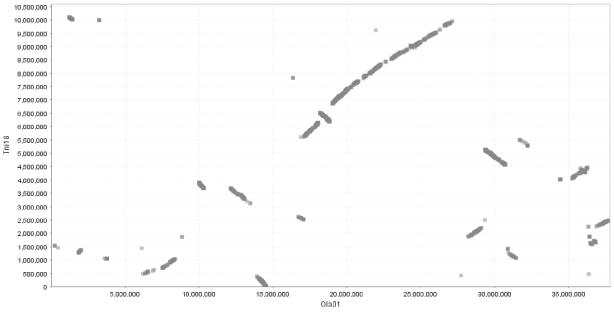
約400万個のヒト EST をヒトゲノムドラフト配列に約１日以内に写像する技術を完成した8)12)．当時としては，類似するソフトウエア sim4 や BLAT と比較して，写像感度をほぼ保持したまま，速度を１桁から３桁改善することに成功した．

その後，様々な種のゲノムが解読され，ゲノムの進化を分析するには，ひとつの種の遺伝子やゲノム断片を異なる種のゲノムにアラインメントすることが必要になった．種を超えてアラインメントする際のマッチ率は70-90%と低い傾向にある．低いマッチ率でも感度のよいアラインメントをするには動的計画法や BLAST の利用も考えられるが，ゲノムへのアラインメントには膨大な時間がかかる．この問題の解決のため BLAT や PatternHunter が普及しているが高速性と高精度を兼ね備えているとは言いがたい．

我々は“multiple-hit mid-spaced seeds” という方式を考案した（特許出願31)．homology ratio > 70% の条件下で，比較ゲノム・アラインメントにおいてしばしば利用される PatterHunter (BLASTZが利用)，BLAT の sensitivity を凌駕することを，塩基変異がランダムに起こるモデルのもと，理論的に示した．それだけに留まらず本方式を実装し，BLAT と比べ主記憶消費量が少ないにもかかわらず，実行スピードが数倍速いことを確認した．完成したソフトウエアALPS 25)は

http://alps.gi.k.u-tokyo.ac.jp/

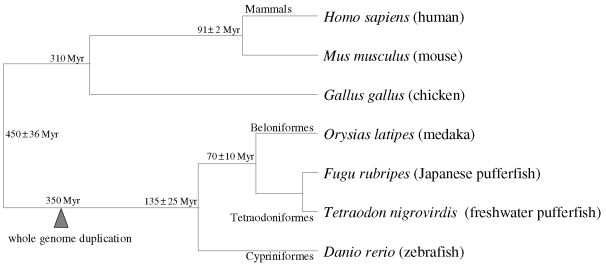
から入手可能．カイコゲノムおよびメダカゲノムへの他の種のESTアラインメントに効果を発揮した（下図参照)．



ミドリフグゲノムの染色体１８番(Tni18, 縦軸) とメダカゲノム染色体1番 (Ola1, 横軸)を ALPS によりアラインメントした結果

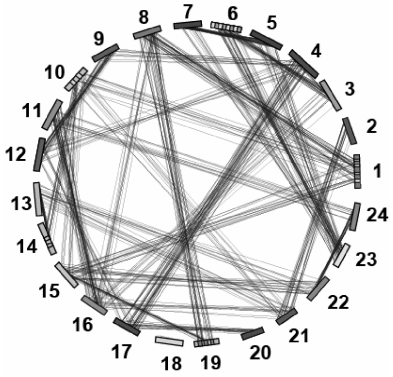
●比較ゲノムとゲノム進化 (平成16年度:中谷洋一郎，曲薇，森下真一)

2001-2005年の間は，ヒト，チンパンジー，マウス，ラット，ドッグなど哺乳類のゲノムの解読が発表され，鳥類ではチキン，魚類ではミドリフグのゲノムが解読された．我々はメダカゲノムを解読した．これら脊椎動物の進化系統樹は下図のように推定されており，解読されたゲノムからゲノムが染色体レベルでどのように組み換えが起こり変化してきたかについて深い関心が寄せられている．



ゲノムが解読済もしくは解読中の哺乳類の進化系統樹

魚類の系統においては，約3.5億年前に全ゲノムが重複されたことがミドリフグのゲノム解読から立証されている．上図においては whole genome duplication と矢印で示された辺りでこの現象が起こったと推定されている．そこで我々もこの議論を確証すべく，メダカゲノム内のパラログ2,075組を用いて染色体間の関係を調べた．その結果，明確に１対１の組をなす染色体が５組同定され，全ゲノム重複が確認された（下図参照）．しかしながら他の染色体については必ずしも対応は明確でなくゲノムの組み換えが起こったことが示唆された．



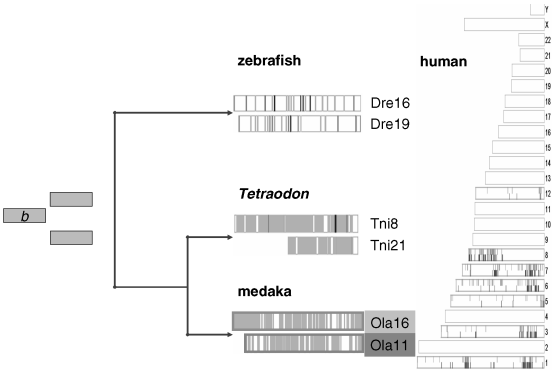
メダカ染色体上に存在するパラログの組を線で結んだ図：染色体が組をなすことが確認された

パラログによる解析では対応が明確にならなかったメダカ染色体がどのように組み替えられてきたかを推定するには，whole genome duplication 以前に分化した生物種のゲノム情報が必要になる．我々はヒトゲノムを利用して whole genome duplication 以前の祖先染色体の状態を推定し，そこからメダカ，ミドリフグ，そしてゼブラフィッシュのゲノムが進化してきた様子を描出することに取り組んだ．まずメダカの予測遺伝子をヒトゲノムにアラインメントし，シンテニーブロックを同定することを試みた．メダカおよびヒト染色体上では進化の過程で数多くの染色体内の逆位が起こっていると考えられるので，メダカゲノム上での遺伝子の順番が必ずしも保存されているわけではない．そのため同じ染色体上にあるか否かという緩やかな判断基準でシンテニーブロックを推定した結果，下図に示すようにまとまったブロックを確認することができた．

メダカゲノム上に存在する予測遺伝子をヒトゲノム上にアラインメントして推定したシンテニーブロック

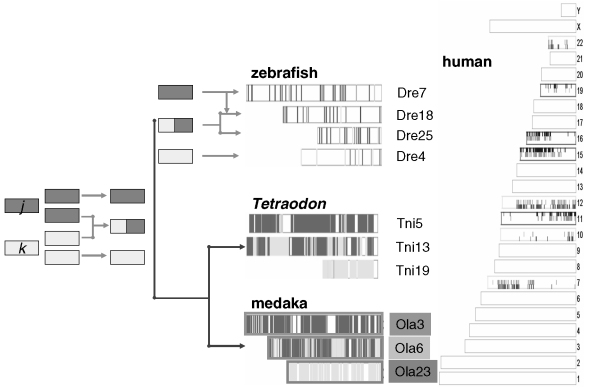
特筆すべき点は，たとえばメダカ染色体５番と７番はパラログ解析から組をなすことが推定できているが，ヒトゲノムへアラインメントした結果，５番７番上の予測遺伝子から推定されたシンテニーブロックが交互にヒトゲノム上において出現していることが確認できたことである．言い換えれば，メダカ内のパラログ解析と，ヒトゲノムを利用したオルソログ解析という独立のアプローチによってwhole genome duplicationが確認でき，シンテニーブロックの正当性も増すこととなった．

このようにシンテニーブロックを多数同定できた．祖先染色体を推定し，そこからの組み換えによるゲノム進化の様子を推定するにはもう一歩踏み込んだ解析が必要であった．すなわちシンテニーブロックの中でどの組が祖先染色体において同じであり whole genome duplication により重複したか否かを推定しなければならない．下図はパラログ解析からも明らかな例であるが，メダカ染色体の１１番と１６番の予測遺伝子をヒトゲノム写像したところ，まさしく同じ染色体上にこれらの遺伝子が並んでいることがわかる．



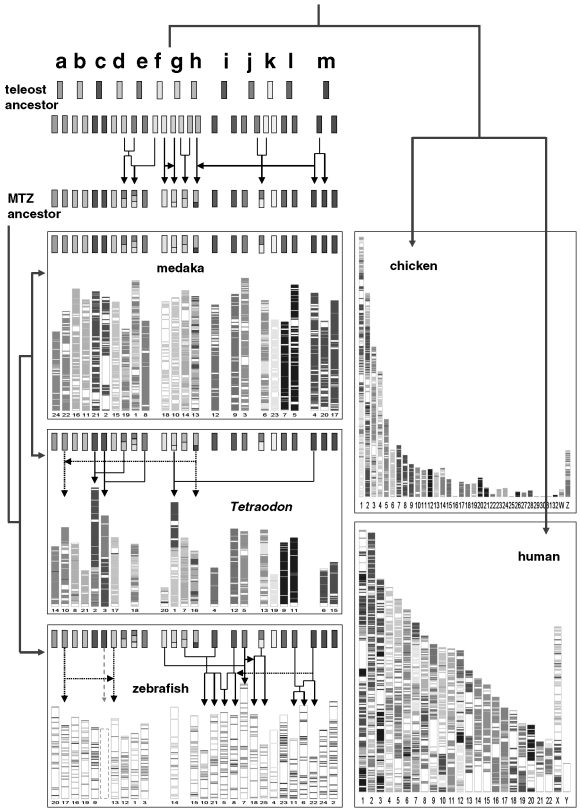
メダカ１１番および１６番染色体上の予測遺伝子をヒトゲノムへアラインメントしてシンテニーブロックを同定した図: パラログ解析からもこれら２つの染色体が先祖染色体をよく保存していることがわかる

下図はより複雑な場合をしめしている．２つの祖先染色体が重複を起こし，各々のコピーが融合した例である．この現象を推測できたのはメダカ染色体３，６，２３番をヒトゲノム写像したとき，３番上の予測遺伝子はヒト染色体１１，１５，１６，１９番に写像されたのに対して，メダカ２３番上の遺伝子は全く異なるヒト染色体７，１０，１２，２２番へと写像され，一方メダカ６番上の遺伝子はどちらのヒト染色体へも写像されたからである．この観測を解釈すれば，メダカ染色体３番と２３番は異なる祖先染色体から由来しているため，ヒトゲノムでも異なった染色体へと写像された．しかしメダカ６番は２つの祖先染色体の融合に由来するためヒト染色体のどちらにも写像されたと考えられる．



メダカ染色体３，６，２３番とヒトゲノム染色体の比較から推定した全ゲノム重複後の２つの祖先染色体 j, k からのゲノム進化の様子

このようにメダカの各染色体の情報を総合することにより，３つの魚ゲノムの進化の様子を描出することができた．ミドリフグの論文でも類似の解析がされているが，曖昧な点が多いことが知られていた．なぜならミドリフグゲノムの染色体被覆率が65%と低かったこと，ミドリフグの系統ではメダカと分化した後に３回の大規模な染色体融合が起こったため，祖先ゲノムの保存状態が崩れている点にあると我々は考えている．一方メダカゲノムの場合は1.3億年前にゼブラフィッシュと分かれた後に大規模な染色体変化を起こしていないため保存度が高く，しかもメダカゲノムの解読可能領域の90%を被覆するだけのゲノム情報を獲得できたため，十分な情報があったことが，正確な解析を可能にした．下図に我々が推定した魚類ゲノムの進化の様子をしめす．



全ゲノム重複以前の祖先染色体 (a - m)の推定とゲノム進化の様子

これらの解析により推定された魚類祖先ゲノムの染色体数は２４である．現在知られている魚類の半数以上についてゲノムの染色体数が２４もしくは２５であり，祖先からの組み替えを鑑みても，今回の推定との整合性が高い．

また今回の推定において１箇所だけ異なるシナリオが考えうる部分がある．それはメダカ染色体１０，１３，１４番の由来である．今回の推定では２つの祖先染色体 g, h が重複し各コピーが融合をおこしたと推定したが，１つの染色体が重複し，１つのコピーが分裂したシナリオも考えられる．２つのシナリオのどちらであるかを判定するには，whole genome duplicationが起こった以前に分化した他生物種のゲノム配列との詳細な比較が必要であろう．

●新規繰返し配列の推定（平成16年度，山田智之）

ゲノムが解読された後には，そのゲノムがどの程度繰返し配列を含んでいるか否かという解析が必要になる．ヒトやマウスのゲノム解析では約半分近くの領域が繰り返し配列により覆われていることもわかった．ヒトやマウスの場合には，既に繰返し配列のデータベースがゲノム解読以前から充実していたため，既知の繰返し配列をゲノムにアラインメントするだけで，ほぼ解析が終了したため技術的には困難でなかった．

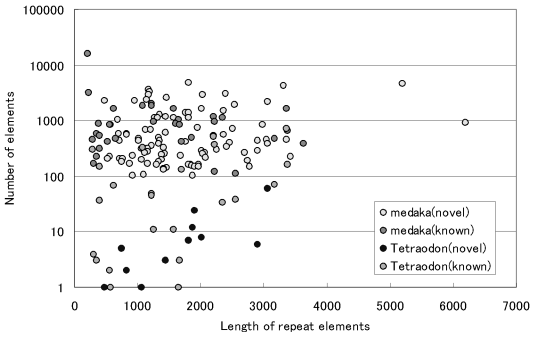
一方，2004年以降に解読されたゲノムの場合，たとえばミドリフグは繰返し配列のデータベースが十分でないため，ゲノム中に含まれる繰返し配列の解析は不十分であり，見落とされている繰返し配列が数多く潜んでいると考えられる．一方，チキンゲノムの解析では，初めて大規模にゲノムから未知の繰返し配列を抽出する解析がなされた．新規繰返し配列の推定は技術的には大変困難であり，しかも膨大な計算時間が必要な問題である．

メダカゲノム解析と平衡して山田智之は独自のアルゴリズムの研究開発に取り組んだ．そのソフトウエアを利用してメダカゲノムから未知の繰返し配列を数多く見出した．下表で示すように，既知の繰返し配列がゲノムに占める割合の総計は 8.3% であったのに対して，山田のソフトウエアが見出した新規繰返し配列はあらたに 9.2%の領域を被覆していることがわかった．

	Masked bases (M b)	Ratio
既知の繰返し配列		
SINEs	5.8	0.8%
LINEs	18.1	2.4%
LTR elements	5.2	0.7%
DNA elements	23.9	3.1%
Small RNA	0.3	0.0%
Satellites	1.1	0.1%
Simple repeats	4.6	0.6%
Low complexity	4.8	0.6%
小計		8.3%
新規繰返し配列	70.2	9.2%
全合計	134.0	17.5%
Total genome size	764.0	

メダカゲノム中の繰返し配列の被覆率に関する統計値：既知の繰返し配列は 8.2% だったのに対して，山田のソフトウエアにより新規に発見された繰返し配列は 9.2% を占めた

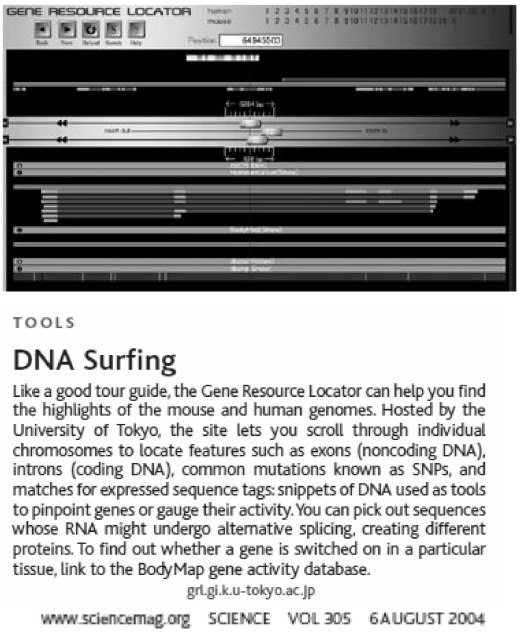
さらに詳細に繰返し配列を調べたところ下図に示すように長さが 1000塩基以上の繰返し配列で頻度分布が1,000 から 10,000 程度の新規繰返し配列の存在も確認され，ミドリフグと共通に保存された新規繰返し配列も見出された．しかしながら哺乳類で存在するような 100万コピーに迫るような繰返し配列は存在せず，最大でも高々20,000コピー程度であった．



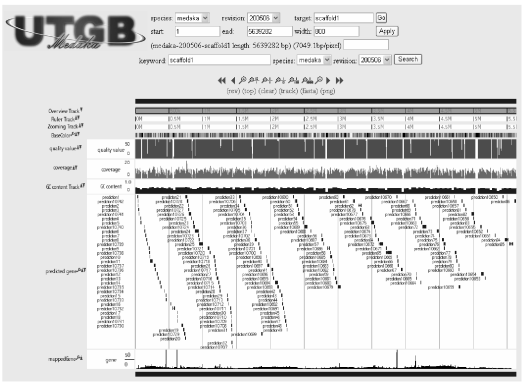
メダカにおける繰返し配列の長さと出現頻度分布

●ゲノム配列をアノテーションするためのゲノムブラウザー　（平成14-16年度: 橋本順之，永安佑希允，笠井康弘，土井晃一郎）

Web サイト GRLでは Flash を用いて動的に内容が変化する GUI を実現した16)．本研究は Science 誌Netwatch の欄でも紹介された35)．



またメダカゲノム解読と平衡して，ゲノムの染色体地図を作成するのにも有効な情報を追加することができるゲノムブラウザーが必要になり新たなシステム UT Genome Browser を研究開発した．このサーバーは公開し，2005年3月からの1年間で約220万ヒットのアクセスが全世界からある．

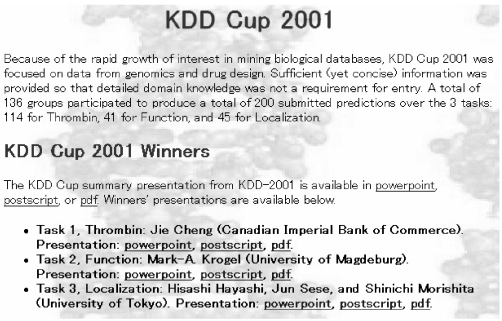


UT Genome Browser

●遺伝子発現量等のデータから機能予測するデータマイニング技術（平成13-16年度: 瀬々潤，林永，森下真一）

ACM SIGKDD 主催の KDD Cup において，遺伝子細胞内局在化部位予測部門で第１位（林永，参加４５チーム中），遺伝子機能予測部門で第３位（瀬々潤，参加４１チーム中）の成果を収めた．

その際，近接点探索に基づくクラス分類アルゴリズムの新手法を林は提案し，森下はその問題の最悪計算量を評価し（NP困難であった），林と森下は現実的に高速に動く分岐限定法を新しく研究開発した14)．



ACM SIGKDD 主催の KDD Cup での優勝チームを紹介した web site

この後注目したのがクラスタリングの限界である．発現量が類似する遺伝子をクラスタリング技術でグループ化した後は，各グループが共通してもつ性質を見つけだす解析がしばしば行われる．しかし直ぐには共通する性質が見つからず，見つかるまで，クラスタリングのパラメータを変更もしくは違ったアルゴリズムで再クラスタリングするような努力が散見される．このような繰返し作業を避けるために，クラスタリング後には各グループが共通の性質をもつことを保証できるアルゴリズム classified clustering を提案した13)．本アルゴリズムの有効性を検証するために，加藤菊也研究室のデータ解析を行い，遺伝子発現量により組織をグループ分けしたとき，各グループがもつ病理的特徴を把握するアプローチを検証した3)．

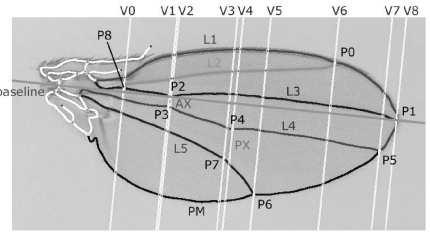
●蛋白質コード領域予測技術（平成13-16年度: 後藤修）

全ゲノム配列が解読された４生物種につき，それぞれ１～４万のイントロン挿入部位を同定し，種に固有な境界シグナルの導出を可能とした．イントロンの長さの分布や通常のギャップをより現実的にモデル化できるようにアルゴリズムを拡張した．二組のテスト用ヒト遺伝子

データセットに対して，開発したプログラム（aln）の予測精度を検証した．ヒト固有の統計情報を使用することと，アルゴリズムの拡張によって，エクソンレベルで4〜5%，遺伝子レベルでは10%以上予測精度が向上することを確かめた．

- ショウジョウバエの翅脈解析ソフトウエア （平成16年度: 中村恵理，初田浩志，森下真一）

相垣研究室で収集しているショウジョウバエ変異体（遺伝子強制発現による）の翅脈画像を処理するソフトウエアの研究に取組んでいる（下図参照）．約3000の変異体から撮影された約20,000枚の画像を処理した結果，翅脈と分岐点を99%以上の精度で正確に自動認識できるまで性能を上げることができた．135個の形態パラメータのデータを取得して，野生株と変異体の間で有意な形態変化が起こっているか調査したところ，変異体の2/3 以上において統計学的に極めて有意な変化が認められた（有意水準 0.002% の両側検定）．



ショウジョウバエの翅脈を認識するソフトウエアによる画像処理

〈国内外での成果の位置づけ〉

- 高感度ゲノミックＰＣＲプライマ設計技術： その後も研究を継続しており，Multiplex genomic PCR への応用を完成させ，その分野の研究では世界をリードしている．
- ヒト／マウス遺伝子阻害のための siRNA配列設計ソフトウエア ： siDirect はsiRNA配列設計では最も利用されている web server となりNucleic Acids Research 誌ウエブサーバー特集号(Vol.32)のトップアクセスペーパー33)．2005年3月からの年間アクセス数は約177万ヒット．
- 5’SAGE解析ソフトウエア： 5’SAGE 法がユニークであるため，同ソフトウエアもユニークである．
- ゲノム配列を解読するWhole Genome Shotgun Assembler の研究開発：Whole Genome Shotgun 方式は新しい生物種のゲノムを解読するたびに新しい問題に直面しそれを解かなければならない．そのため約１０個の同種のソフトウエアが存在するが比較するのも困難である．我々のアセンブラはメダカとカイコ固有の問題にチューンされているが将来的には普遍的な問題が解けるように改善したい．
- メダカゲノム概要配列決定： 平成１６年度末には間に合わなかったが，平成１７年度に入って完成した．
- 5’SAGE法を使った遺伝子推定：他に類を見ない方法論である．
- 比較ゲノムのための高速アラインメントソフトウエア ：同種のソフトウエア Blastz や PatternHunter と同程度の性能は持ち，フリーソフトウエアとして配っているが，残念ながら知名度が劣っている．
- 比較ゲノムとゲノム進化：魚類ゲノムの進化を明らかにできた点でユニークな研究となった．
- 新規繰返し配列の推定： チキンゲノムから繰返し配列を推定したソフトウエアと山田方式の２つしか存在しない．比較を通じて優位性を明らかにする段階．
- ゲノム配列をアノテーションするためのゲノムブラウザー：残念ながら NCBI, Ensembl, UCSC と比べて知名度は劣る．開発した UT Genome Browser は Source Forge から無料で配布．メダカゲノムサーバーは2005年3月からの年間アクセス数が約220万ヒット．本来であれば，この分野でも世界をリードする web server を構築したかったが，海外の有力サイトに対抗するのが現状は残念である．
- 遺伝子発現量等のデータから機能予測するデータマイニング技術： ACM KDD Cup での優勝により我国のデータマイニング技術が世界的なレベルにあることはアピールできた．
- ショウジョウバエの翅脈解析ソフトウエア：他には類の無いユニークな研究である．

〈達成できなかったこと，予想外の困難，その理由〉

特定領域研究に参加してゆく中で様々な共同研究の機会に恵まれて，研究期間内に全テーマを達成するのは困難であった．

メダカゲノム解析を平成１６年度末までに終了できず，平成１７年度末まで取り組んだ点が最も計画外のことであったが，最終的には当初の目標を達成できた．

ゲノムブラウザーについては当初は世界的な web server にすることを目標にしていたが，海外の大きなセンターを背景にした web site である NCBI, Ensembl, UCSC と伍してゆくのは困難であった．大学の少人数の研究室で対抗するのは難しいことを痛感した．

〈今後の課題〉

ショウジョウバエの翅脈の表現型解析と siRNA 配列による遺伝子阻害の研究は平成１７年度から始まった特定領域研究「生命システム」のなかで継続して研究することができるようになり継続している．表現型を定量化することにより遺伝子機能を推定することが今後の目標である．

メダカゲノム情報は各国からの要望も大きいため Ensembl にデータを渡して UT Genome Browser と相互接続して広く公開することとなった．メダカゲノムは概要配列の段階で，多数の小さなギャップを埋め，あたらしいゲノムマーカを使って染色体の被覆率を上げることが大事であり継続したい．

ゲノム情報，遺伝子発現量情報，表現型情報などのマルチモーダル情報を巧みに組合せて実験計画を立てることができるシステムへのニーズが高いことがわかった．しかし現実には，個々のデータ処理は五月雨的かつ可及的すみやかな対応が求められるため，どうしても付け焼刃のスクリプトプログラミングで対応する場面が多かった．今後は．マルチモーダル情報を自動的に組合せて実験計画を立てることができる統合データベースシステムを，時間をかけながらじっくりと構築することが必要と考えている．

〈研究期間の全成果公表リスト〉

1） 論文／プロシーディング（査読付きのものに限る）

1) 0601311156 Tomoyuki Yamada and Shinichi Morishita. Accelerated off-target search algorithm for siRNA.

Bioinformatics, 21(8):1316-1324, (Jan., 2005)
2) 0601311201 Yasuhiko Kasai, Shin-ichi Hashimoto, Tomoyuki Yamada, Jun Sese, Sumio Sugano, Kouji Matsushima, and Shinichi Morishita. 5'SAGE: 5'-end Serial Analysis of Gene Expression database. Nucl. Acids Res. Database Issue 33: D550-D552 (Jan., 2005).
3) 0408192354: Jun Sese, Yukinori Kurokawa, Morito Monden, Kikuya Kato and Shinichi Morishita. Constrained Clusters of Gene Expression Profiles with Pathological Features, Bioinformatics, 20: 3137 - 3145, (2004)
4) 0409091457: Shin-ichi Hashimoto, Yutaka Suzuki, Yasuhiro Kasai, Kei Morohoshi, Tomoyuki Yamada, Jun Sese, Shinichi Morishita, Sumio Sugano, Kouji Matsushima. 5'-end SAGE for the analysis of transcriptional start sites. Nature Biotechnology 22, 1146 - 1149 (2004)
5) 0408192359: Yuki Naito, Tomoyuki Yamada, Kumiko Ui-Tei, Shinichi Morishita, and Kaoru Saigo. siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. Nucl. Acids Res. 32: W124-129 (2004)
6) 0408192336: Tomoyuki Yamada and Shinichi Morishita. Computing Highly Specific and Noise Tolerant Oligomers Efficiently, Journal of Bioinformatics and Computational Biology, Vol. 2, No. 1, pp21-46 (2004)
7) 0408192332: Kazuei Mita, Masahiro Kasahara, Shin Sasaki, Yukinobu Nagayasu, Tomoyuki Yamada, Hiroyuki Kanamori, Nobukazu Namiki, Masanari Kitagawa, Hidetoshi Yamashita, Yuji Yasukochi, Keiko Kadono-Okuda, KimikoYamamoto, Masahiro Ajimura, Gopalapillai Ravikumar, Michihiko Shimomura,Yoshiaki Nagamura, Tadasu Shin-i, Hiroaki Abe, Toru Shimada, Shinichi Morishita, and Takuji Sasaki. The Genome Sequence of Silkworm, Bombyx mori. DNA RESEARCH Vol.11, No.1, pp.27-35 (2004)
8) 0404071927 Jun Ogasawara and Shinichi Morishita: Fast and Sensitive Algorithm for Aligning ESTs to Human Genome.Journal of Bioinformatics and Computational Biology, Vol.1, No.2, 363-386, (2003)
9) 0404071930 Tomoyuki Yamada and Shinichi Morishita: Computing Highly Specific and Mismatch Tolerant Oligomers Efficiently, Proc. of Second IEEE Computer Society Bioinformatics Conference (CSB2003), Stanford University, Palo Alto, CA August 11-14, 316-325, (2003)
10) 0404071942 Shinichi Morishita and Asao Fujiyama: Body Expression Map of the Human Genome, Encyclopedia of Molecular Cell Biology and Molecular Medicine, 2nd Edition, Wiley-Vch, 75-85, (2003)
11) 0404072009 Koichiro Doi, Jing Li, Tao Jiang: Minimum Recombiant Haplotype Configuration on Tree Pedigrees. WABI 2003: 339-353 (2003)
12) 304261338: Jun Ogasawara and Shinichi Morishita: Fast and Sensitive Algorithm for Aligning ESTs to Human Genome. Proc. of First IEEE Computer Society Bioinformatics Conference, Stanford University, Palo Alto, CA, 43-53, (2002).
13) 304261344: Jun Sese and Shinichi Morishita: Answering the Most Correlated N Association Rules Efficiently. Proc. of 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, (2002).
14) 304261354: Jie Cheng, Christos Hatzis, Hisashi Hayashi, Mark-A. Krogel, Shinichi Morishita , David Page, and Jun Sese: KDD Cup 2001 Report. ACM SIGKDD Explorations, Volume 3, Issue 2, 47-64 (2002).
15) 304261403: Shinichi Morishita: Computing Optimal Hypotheses Efficiently for Boosting. Progress in Discovery Science, 2002, Springer, 471-481 (2002).
16) 202271440: Toshihiko Honkura, Jun Ogasawara, Tomoyuki Yamada and Shinichi Morishita:The Gene Resource Locator: gene locus maps for transcriptome analysis. Nucleic Acids Research, 2002, Vol. 30, No. 1 221-225 (2002).
17) 0202271459: Yasuhiko Morimoto, Hiromu Ishii, and Shinichi Morishita. Efficient Construction of Regression Trees with Range and Region Splitting. Machine Learning, Kluwer Academic, 45, pages, 235-259 (2001)
18) 0110291804: Jun Sese, Hitoshi Nikaidou, Shoko Kawamoto, Yuichi Minesaki, Shinichi Morishita, and Kousaku Okubo.BodyMap incorporated PCR-based expression profiling data and a gene ranking system. Nucleic Acids Res. 29: 156-158. (2001)
19) 0110291801: Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita and Takeshi Tokuyama. Data Mining with Optimized Two-Dimensional Association Rules. ACM Transactions on Database Systems (TODS), Volume 26 , Issue 2, pp. 179 - 213, (2001).

2）データベース／ソフトウェア

20) メダカゲノムブラウザー UTGB (Medaka)
http://medaka.utgenome.org/
21) siRNA配列設計 siDirect
http://design.rnai.jp/
22) dsRNA 配列設計検証 dsCheck
http://dscheck.rnai.jp/
23) 5'末端SAGEデータベース 5'SAGE
http://5sage.gi.k.u-tokyo.ac.jp/
24) 出芽酵母遺伝子破壊株イメージ処理データベース SCMD
http://scmd.gi.k.u-tokyo.ac.jp/
25) 遺伝子配列のヒトゲノムへのアラインメント ALPS
http://alps.gi.k.u-tokyo.ac.jp/
26) Genomic PCR プライマー設計ツール GenomeSURF
http://surf.gi.k.u-tokyo.ac.jp/
27) ヒト血液系 SAGE 発現プロファイル
http://bloodsage.gi.k.u-tokyo.ac.jp/
28) ヒトゲノムブラウザー GRL
http://grl.gi.k.u-tokyo.ac.jp/

3）特許など

29) siRNA 配列設計法に関する特許 2 件
(RNAi社より出願)
30) オリゴ設計法に関する特許 1 件
（東大ＴＬＯより出願）
31) 遺伝子アラインメント法 1 件
（東大ＴＬＯより出願）
32) アソシエーションルールによるクラス分類 1 件
（東大ＴＬＯより出願）

4）その他顕著なもの

33) 2001 年8 月 ACM SIGKDD 主催 KDD Cup 優勝
http://www.cs.wisc.edu/~dpage/kddcup2001/

34) 2004年8月13日　siRNA設計サイト siDirect がNucleic Acids Research 誌ウエブサーバー特集号(Vol.32)のトップアクセスペーパー
http://www3.oup.co.uk/jnls/list/nar/special/11/default.html

35) 2004年8月9日　サイエンス誌ネットウォッチ(Science Vol 305, 6 August 2004) でヒト/マウスゲノムブラウザー Gene Resource Locator (GRL) が紹介
http://www.sciencemag.org/cgi/reprint/305/5685/759e.pdf