Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

頻出パターン解析による重複遺伝子の同定手法

小中佳子[†1] 福崎睦美[†1] 吉田真明[†2]
清水健太郎[†3] 小倉淳[†2] 瀬々潤[†1]

生物が新規機能を獲得する上で，遺伝子重複は重要なイベントのひとつである．遺伝子重複により，いずれかの遺伝子に変化が起きても細胞中に変化が現れることなく，その後の変異を蓄積して行くことが出来るからである．その一方で，重複した遺伝子は互いに近い配列を持っているため，単に遺伝子配列を読んだだけでは，同一の遺伝子のシークエンスエラーなのか，本当に複数の遺伝子が存在しているかの区別を付けるのが難しい．そこで，今まで重複が起こった遺伝子群を見つけるには，ゲノム領域の塩基配列を決定するのが一般的である．しかし，ゲノム配列の決定は現在においても非常にコストが高く容易に決定できるものではなく，特に真核生物では限定的な種でしかゲノム配列が決定できていない．本研究では，次世代シークエンサを用いてmRNA を大量に読んだデータを用い，シークエンスエラーによる変異と，重複遺伝子による変異を区別する手法を提案する．提案手法は，データマイニングの相関ルールを応用した手法であり，提案手法を用いることで，重複遺伝子の各配列と，近縁種から変異の起こった位置を計算する事が可能である．本研究では，このアルゴリズムを軟体動物 4 種の遺伝子領域を次世代シークエンサで読んだ配列に適用する事で，各種から重複遺伝子候補を抽出することに成功した．抽出された重複遺伝子にはヒストンクラスタが含まれており，本アルゴリズムが妥当である事が示された．また，zinc finger 遺伝子の重複による機能分化も観測された．

Gene duplication detection with frequent pattern analysis

YOSHIKO KONAKA,[†1] MUTSUMI FUKUZAKI ,[†1]
MASAAKI YOSHIDA ,[†2] KENTARO SHIMIZU ,[†3]
ATSUSHI OGURA [†2] and JUN SESE [†1]

Gene duplication is one of the important events for the gain-of-function. The reason is that mutation of one of the duplicated genes will not affect on the function of cells because the alternative duplicated gene will work and can keep the cellular function. On the other hand, it is difficult to determine the duplicated genes from gene sequences in non-model species because of the high similarities of gene sequences between duplicated genes. Therefore, most of known duplicated genes have been found in species whose whole genome sequences are known.

In this study, to avoid high cost and time consuming whole genome sequencing, we propose techniques to determine duplicate genes by using large amount of mRNA sequences observed by next-generation sequencer and their mutation positions. We applied frequent pattern mining technique for detecting mutated regions, and the method allows us to compute gene sequence of the duplicated genes and mutated positions from closely related species. In this paper, we applied the algorithm for four different mollusks data observed by next-generation sequencers, and successfully predicted more than hundred duplicated genes, including zinc finger protein whose both sequences and functions are diverged from related species.

1. はじめに

進化の過程でどのように機能獲得が起こるのかを知る事は，進化を知る上で重要な課題である．機能獲得で一つの重要なイベントは，機能の重複であり，重複した機能は一方がその機能を欠失しても，もう一方が働くことから既存の機能を基にして，新たな機能を獲得したり，有している機能に関し潜在的な別の使い方をする事が可能となる．このような現象の最たる物は，全ゲノム重複や遺伝子重複であり，例えば酵母においては全ゲノム重複が起きた後，多くの遺伝子は活性を失うが，457 の遺伝子は未だに発現をしている[1)] ことや，遺伝子制御ネットワークは遺伝子の重複が重要である[2)] といった研究結果が発表されている．

その一方で，ゲノムや遺伝子の重複を見つけることは必ずしも容易ではない．特に進化的に最近起こった重複の場合，変異の入る量が少なく，ある配列が複数の領域から採取されたものなのか，単一の領域から採取されたが，配列決定時のミスにより，複数種類の配列が存在するように見えてしまうのかの区別を付けることが必ずしも容易ではないためである．このため，重複遺伝子の探索は，全ゲノム配列を決定し，その後，遺伝子間の配列相同性を調べる手法が取られることが多かった．

本研究では，この全ゲノム配列決定を避け，遺伝子領域のみのシークエンスから遺伝子重

†1 お茶の水女子大学 大学院人間文化創成科学研究科
Graduate School of Humanities and Sciences, Ochanomizu University.
†2 お茶の水女子大学 アカデミックプロダクション
Ochanomizu University Academic Production.
†3 チューリッヒ大学
University of Zurich

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

複が起こっていることを推定する手法を導入する．より具体的には，次世代シークエンサで読まれた大量のリード配列に対し，近縁種のゲノム配列を参照する事で，リードと近縁種ゲノムの間にある変異位置を同定する．そして，その変異位置が共起するリードを調べてグループ化する事で，重複遺伝子の同定を行う．更に本手法を，軟体動物 4 種から得られた RNA-seq 配列に適用し，重複遺伝子を同定した．

2. 手　　法

本研究の目標は重複遺伝子を次世代シークエンサで得られた大量配列から同定することである．特に，ゲノム配列は読まれていないが，EST を大量に読んでいるケースを想定している．同定に当たり，なるべく近い種の遺伝子配列を参照する．この参照する種は必ずしも近縁である必要はない．後に述べる本研究の実験結果では，イカやタコといった軟体動物・頭足類の配列から重複遺伝子候補を見つけるが，この際に参照する配列は軟体動物・腹足類のカサガイを利用して発見を行っている．参照する配列が近縁であればあるほど，より確実に遺伝子の重複を同定することが可能である．

本章ではまず, 手法の概要と困難となる点を示し, 次に重複遺伝子発見の詳細を議論する．

2.1 手法の概要

本節では手法の概要を示し，重複遺伝子の同定において困難となる点を明らかにする．手法の概観を図 1 に示す．

本手法の入力は二種類あり，図 1(A) 次世代シークエンサで読まれた RNA-seq 配列 (以下リード配列)，図 1(B) 近縁種の遺伝子配列，である．実験では次世代シークエンサとして 400bp 弱を読むことができる Roche 社の 454 を利用しているが，導入する手法はシークエンサの特性によるものではなく，より長い配列が読めるサンガー法で大量に配列を読んだ場合や，より短い配列を大量に読める Illumina 社の Solexa や，ABI 社の SOLiD で読まれた配列であっても，同様のアルゴリズムで重複遺伝子の同定が可能である．本手法をこれらのデータに適用する事により，(1) 重複遺伝子に由来すると考えられる配列群と (2) それらの配列が重複遺伝子由来と判定した塩基配列群を得ることが可能である．重複遺伝子と見込まれる配列群に加え，判定由来の塩基を明示することで，擬陽性の検査を生物学的知識を用い行うことが可能であること，及び，次世代シークエンサ特有のシークエンスエラー由来の擬陽性であるかどうかも確認が可能である．

入力に対し，各リードが既知のどの遺伝子のどの領域に対応する可能性があるのかを知るため，配列アラインメント手法を用いて，リード配列を近縁種の遺伝子配列に対応付ける (図 1(C))．配列アラインメント手法としては，BLAST[4] や BLAT[5]，あるいは次世代シークエンサ用に開発されている手法[6),7)] など，既存の手法を用いることができる．どの手法を選ぶかは，リード配列の種とアラインメント対象の種の進化距離に依存する．一般に進化距離が遠いほど，リードと遺伝子の対応に関して不確実性が増すため，問題は難しくなる．

アラインメント結果を基に，重複遺伝子の同定を行う (図 1(D))．同定に際し，頻出パターン解析を利用したクラスタリング手法を提案する．詳細は次節で述べる．本手法により，各近縁種の遺伝子から見て，(1) 重複遺伝子由来と考えられるリード群が存在するか否か，(2) 存在するとすれば，どのような配列であるか，(3) その配列群が重複遺伝子由来と判定できる理由，を得ることができる．重複遺伝子同定において難しいことは，対象種の遺伝子とリード配列の間で塩基置換が認められた場合に，それがシークエンスエラーなのか，SNP なのか，重複遺伝子が存在するために起こったのかを判定することである．本研究では，次世代シークエンサがリードを数多く読むことができる利点を利用し，1 本のリードの中で，複数箇所で共通の塩基置換が起こり，かつ，そのような配列が複数存在する場合に，シークエンスエラーでも，SNP でも無く，重複遺伝子である可能性が高いため，このモデルを利用して重複遺伝子の同定を行う．

最後に，本手法では対象の近縁種の遺伝子に重複あるいは，強いホモログが存在する場合にも，重複遺伝子として同定してしまう可能性が高い．このような擬陽性を取り除く．

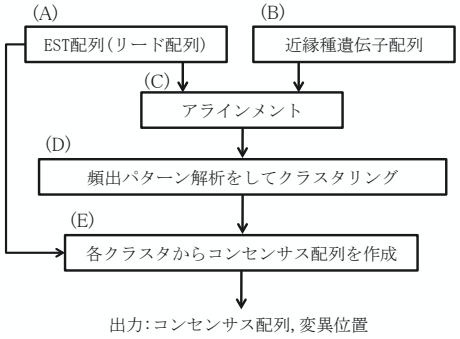

図 1　手法の概観

2.2 変異の頻出パターン抽出

本節では，近縁種の遺伝子 1 つと，それにアラインメントできたリード配列群を入力と

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

し，対象種において遺伝子重複が起こったか否か，起こっている場合は，対処種においてどのような遺伝子配列となっているかを得る．

ある種において重複遺伝子が観測される場合とは，祖先のゲノム配列から，特定領域の重複が起き，その領域に遺伝子が含まれており，かつ，もとの遺伝子と重複後の遺伝子の両方が発現している場合である．このような場合，重複前の遺伝子と重複後の遺伝子は，重複後の進化の過程では独立に塩基置換が起き，結果として，重複遺伝子はほとんど同一の配列を持ちながら，微妙に異なる配列を有することになる．

このような遺伝子群から観測されたリードを，近縁種の遺伝子配列と比較してみると，近縁種との間に多少の塩基置換が観測されることとなる．

ところで，近縁遺伝子の配列とリード配列がどのような関係にある場合に，遺伝子重複と考えられるかを調べる．近縁種の遺伝子とリード配列が異なる場合，次の 3 種類の変異が考えられる．(1) シークエンサの出したリード配列にエラーがある場合，(2) 近縁種からの変異が起こっている場合，(3)SNPs による個体差が起こっている場合，(スプライスバリアントの変化等によって起こった変異は，単一塩基の置換としては観測されず，重複遺伝子の同定には関与しないため，ここでは局所的な変異のみを考える.)

ここでは重複遺伝子を観測するため，重複遺伝子の近縁種からの変異の入り方に着目をする．図 2 に遺伝子重複と，それらに入る変異の関係の模式図を示した．遺伝子 A が重複して A', A" になった場合，それらは独立して変異が入るので，異なった場所に変異が入る可能性が高い．一方，我々が読むリードは，これらの遺伝子の区別無く読まれるため（図 3)，A' 由来のリードも A" 由来のリードも，同一の近縁種遺伝子にアラインメントされる．この複数のリードを見比べると，A' 由来のリードは A-A' 間で入った変異と同様の変異が観測され，A" 由来のリードは A-A" 間で入った変異と同様の変異が観測される．これらの変異の位置が異なることを利用して，ある近縁種の遺伝子毎に，複数のパターンで変異が観測されれば，その遺伝子が対象種に進化する上で重複した可能性があることを示唆することになる．さらに，この変異が複数のリードに跨って観測されれば，より重複である確度が高くなる．

2.2.1 頻出パターン抽出

本節では頻出パターン抽出と重複遺伝子判定の対応付けをおこない，重複遺伝子判定の手法を導入する．

頻出パターン抽出において，データベース中に含まれるトランザクション集合の要素には，トランザクション ID が割り振られており，各トランザクションはアイテム集合を有す

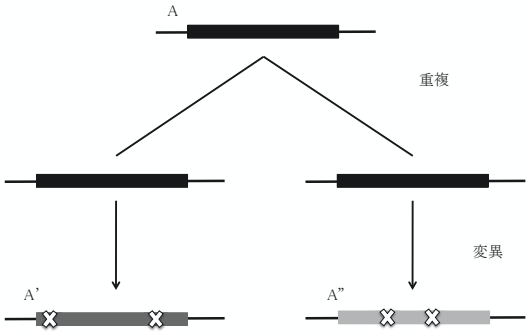

図 2 遺伝子重複

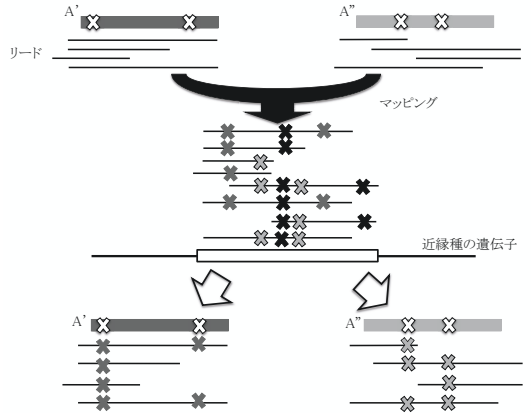

図 3 重複遺伝子発見の概要

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	A	A	G	T	G	G	T	C	C	C	T	G	T	A	G	C	C	C	T	T	T	A	G	G	T
r1	T	T	G	T	G	G	A	G	T	A	A	G	T	A	T	C	G	C	G	T	T	G			
r2				T	G	G	T	G	T	T	A	G	T	A	G	C	G	C	G	T	T	G	G		
r3						G	A	G	T	A	A	G	T	A	T	C	G	C	G	T	T	G	G	G	C
r4			G	T	G	G	A	G	T	A	A	G	T	A	T	C	G	C	G							
r5	T	T	G	A	G	G	T	G	T	C	A	C	T	A	G	C	G	G	G	T	T	G	G		
r6				A	G	G	T	G	T	C	A	C	T	A	G	C	G	G	G	T	T				

図 4 アラインメント例

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

表 1 変異位置とリードを表すデータベース．

TID	リード集合
1	r1,r5
2	r1,r5
4	r5,r6
7	r1,r3,r4
8	r1,r2,r3,r4,r5,r6
9	r1,r2,r3,r4,r5,r6
10	r1,r3,r4
11	r1,r2,r3,r4,r5,r6
12	r5,r6
15	r1,r3,r4
17	r1,r2,r3,r4,r5,r6
18	r5,r6
19	r1,r2,r3,r4,r5,r6
22	r1,r2,r3,r5
25	r3

表 2 1-アイテム集合に相当する 1 変異を網羅した表．

リード集合	共通変異数
{r1}	11
{r2}	6
{r3}	10
{r4}	8
{r5}	11
{r6}	8

表 3 2-アイテム集合に相当する 2 個の変異が共起しているリードと，共通する変異数．

リード集合	共通変異数
{r1,r2}	6
{r1,r3}	9
{r1,r4}	8
{r1,r5}	8
{r1,r6}	5
{r2,r3}	6
⋮	⋮

る．アイテム集合は，0 個以上のアイテムの集合である．頻出パターン抽出は，このデータベースから一定回数以上（最小サポート数以上）現れるアイテム集合を列挙する問題である．

本研究では，近縁種遺伝子の各塩基をトランザクションに対応させ，近縁種遺伝子から変異が認められた場合，変異の位置に相当するトランザクションがリード番号に相当するアイテムを有すると考える．図 4 に近縁種にアラインメントした例を示す．一行目は便宜上の塩基番号を示し，二行目に近縁種の遺伝子配列を示している．三行目以降各行はリードを示している．最も左の行には，リード番号を示した．

いずれのリードも，近縁種遺伝子と対応付けられてはいるものの，変異が認められる．データベースのトランザクションの形式で変異を表したものが表 1 となる．各トランザクション ID(TID) は，塩基番号に対応しており，TID が 1 のトランザクションは，図 4 より近縁遺伝子では A であるが，今回の実験では r1, r5 の 2 回のリードで観測されており，いずれも塩基は T である．これより，TID1 は，塩基 1 で変異するリードの集合 r1, r5 を有している．このリード集合は頻出パターン解析のアイテムに相当する．また，表 1 では，スペースの都合によりデータベース中には塩基置換の認められなかった（アイテムがひとつも無い）トランザクションは予め除外してある．

表 1 より，頻出パターンを発見できるアルゴリズムの一つである Apriori[8] と同様の手法を用い，頻出する変異を見つける．あるアイテム集合が頻出か否かを認定するため，最小サポート c を予めユーザが定める．サポートとは全トランザクションに対して着目するアイテム集合を含むトランザクションの数の事であり，Apriori では，データベースより最小サポート以上の割合を持つアイテム集合を列挙する．本提案アルゴリズムでも同様に，特定の遺伝子に着目した場合の全変異位置に対し，一定割合 c 以上の変異を共有するリード集合を求める．

より具体的には，まず 1 個の変異に着目した場合のサポートを計測する．表 2 に表 1 から求めた例を示す．各リード毎に変異の個数が計測される．もし，この時点で最小サポートを切るリードがあれば表から削除される．

この表を基に，2 個のリードの組み合わせを考える．表 3 には 2 個のリードの組み合わせについて計測した例を示している．この際も同様に最小サポートを切るリードがあれば表から削除される．以上の操作を繰り返し，最小サポートを上回るアイテム集合が無くなった時点で，本計算は終了する．

求まった頻出パターンは，重複遺伝子候補と考えられる．なぜなら，そのリード群は近縁遺伝子から見て共通の変異パターンを示しており，かつ，そのリード群に入らないリードが多数存在するとすれば，遺伝子を読んだ種には二種類以上の遺伝子の存在が示唆されるからである．

頻出パターンを列挙する際はアイテム集合をすべて列挙するが，本研究の目的を考えると，興味が有るのは，アイテム集合のスーパーセットが頻出でない，つまり，あるリード群に，新たなリードを加えると変異数がサポートを必ず切るような集合である．新たなリードが加わっても頻出なのであれば，本研究ではそのリードは加えた集合を考えたほうがより同一の遺伝子から多くのリードが読まれている状態になるため，好ましい．このような問題は頻出パターンの文脈では，極大アイテム集合として扱われている．本研究では，後述するように極大アイテム集合が必ずしもそのまま重複遺伝子に対応せず，本問題固有のヒューリスティクスを入れる必要があるため，同様のアプローチは取らず，通常どおり頻出パターンを列挙する．

2.2.2 頻出パターンからのリード集合選択

本節では，求まった頻出パターンから，重複遺伝子候補に相当するリード集合を選択する．これにより，各遺伝子から何種類の重複遺伝子候補を考えることができるかが分かる．

リード集合の選択に際し，最も長いリード集合（リード数の多い集合）に着目し，この集合数により，場合分けをする．これは，最も長いリード集合がひとつの場合はそのリード集合を重複遺伝子の第 1 候補と考え，それらのリードから重複遺伝子配列を構成できるが，複

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

数ある場合にはそのすべてから重複遺伝子の候補を生成することが妥当であるか，それらの間には何らかの重複があり，リードエラー，あるいは，個体差に寄る塩基置換によって引き起こされたノイズであるのかを判定する必要があるためである．

(1)　最長のリード集合が 1 つの場合

最も長いリード集合が唯一に決まる場合は，そのリード集合をひとつのグループとみなす．その後，次に長いリード集合中で，上記グループとは最も離れているリード集合を探し，再帰的に同様の操作を行う．リード集合間の類似度は，二つのリード集合と R_1, R_2 とすると，

$$D(R_1, R_2) = \frac{r \times \min\{|R_1|, |R_2|\} + \max\{|R_1|, |R_2|\}}{|R_1 \cup R_2|}$$

で定義する．$|\cdot|$ は要素数である．R_1 と R_2 の間に重複が多いほど，$D(R_1, R_2)$ は大きくなる．

(2)　最長のリード集合が 2 つの場合

この場合は, その 2 つのリード集合が, 本当は 1 つの遺伝子に由来し，計測ノイズや SNPs により現れた集合か，本当に二つの遺伝子に由来するかのいずれかである．この判定のために，上記で定義した $D(R_1, R_2)$ を利用する．2 つのリード集合を R_1, R_2 とし，ユーザが予め指定する閾値を d とすると，$D(R_1, R_2) > d$ の場合に 2 つのグループであるとみなす．この場合には，2 つの重複遺伝子から採取されたリード群が観測されたとして，終了する．$D(R_1, R_2) \leq d$ の場合，2 つの集合は同一の遺伝子から観測された可能性が高いため，$R_1 \cup R_2$ なる新たな集合を作成し，(1) を実行する．

(3)　最長のリード集合が 3 つ以上の場合

全リード集合間の類似度を，$D(R_1, R_2)$ を用いて計測し，クラスタリングの最短類似度法と同様の手法で集合をまとめていく．つまり，まず，最も近いリード集合 R_i, R_j を見つけ，$R_i \cup R_j$ なる集合を構成する．次に，現在存在するリード集合間で最も近いリード集合を見つけ併合する，という操作を繰り返す．この操作は，最も近いリード集合間の類似度が (2) で導入した d 以下になった時点で終了する．この操作により，ひとつの集合になった場合には (1) と同様の操作を行い，二つ以上になった場合には，各集合がそれぞれ異なる遺伝子由来であると考えて終了する．

以上のようにクラスタリングを行う．しかし，このままではうまくグループ分けができない場合がある．それは，全体のミスマッチ位置が多く，かつ重複遺伝子毎にグループ分けをする際に不要なミスマッチがある場合である．このような場合では，全体のミスマッチ位置の数が多くなり，しきい値の割合を満たさなくなる場合や，重複遺伝子として関わりのないリードもグループの中に含まれてしまい，結果として間違ったクラスタリングをしてしまう場合がある．そこで本研究では，重複遺伝子を見つける際に不要なミスマッチ位置を削除できるような，データベースのフィルタリング手法を提案する．

リードのエラーや SNPs に起因するエラーを減少させるため，データベース構成後，以下の 3 つの場合の変異位置を TID から削除する．

(a)　アラインメントされたリード全てに共通したミスマッチ位置がある場合

(b)　ミスマッチ位置をもつリードが 1 本だけの場合

(c)　アラインメント領域に含まれるリードすべてがミスマッチ位置をもつ場合

これにより, 本来の計算に必要ない変異位置を削除することができ，グループ分類の精度が向上するとともに，データベースが小さくなるので実行時間の短縮が見込める．

2.3　代表配列作成手法

2.2.2 節で作成したリード集合は，互いに同一の塩基置換を共有しているため，ほぼ同一の配列であることが期待されるが，シークエンスエラーなどから実際には多少配列が異なっている．本節では，これらのリード集合から，これらのリード集合を表すコンセンサス配列を求める手法について述べる．各リード配列群に対し，以下の手順を行う．

(1)　リード配列の向きをそろえる：この操作は次世代シークエンサー特有の内容となるが，サンガー法とは異なりリードの向きが分からない為，向きをそろえる必要がある．今回は近縁種の遺伝子にアラインメントできるリードのみを利用しているため，近縁種の遺伝子にアラインメントされた向きにそろえる．

(2)　欠失，挿入位置の同定と補正：近縁種の遺伝子との配列を比較する事で，欠失，挿入位置を同定する．重複後の遺伝子が機能していると仮定した場合，フレームシフトを考えると，特にコーディングリージョンでは 1 塩基の欠失や挿入が起こる可能性が低いため，1,2 塩基の欠失や挿入は，リードのミスであると考える．特に Roche 社の 454 では，このような 1 塩基多く観測する現象が知られており，配列を正しく保つためには，有用な補正である．また，1,2 塩基であっても，多くのリードで同様の変異が見られる場合，及び，3 塩基以上の変異は，実際に起きていると見なし変異を残してコンセンサス配列を求める．以上の操作は，近縁種遺伝子にアラインメントされた領域において行う．

(3)　アラインメント領域外の推定：参照した近縁種遺伝子にアラインメントできない領域であっても，その近縁種遺伝子に変異が大きかったり，リードを読んだ種に変異が大きい場合，塩基配列が読めている可能性が有る．この領域の配列を推定するため，

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

前項の配列位置補正位置を信用し，各位置で現れる最も多い塩基をコンセンサス配列として決定する．同一の位置に複数の塩基が同等に現れる場合，コンセンサス配列は”N” とする．

本節で導入した手法により，各リード集合からコンセンサス配列が得られ，リードを読んだ対象種で重複した遺伝子の配列を決定することができる．

3. 実行結果と考察

前章までに導入した重複遺伝子発見手法により，重複遺伝子が同定出来るか否かを調べる為，ゲノム未知の軟体動物のヒメイカ，ヤリイカ，ホタテガイ，オウムガイから mRNA を採取し，ノーマライズ処理をした後，Roche 社 454 Titanium で配列の決定を行った．

3.1 実行データと結果

4 種の軟体動物と，近縁種としてカサガイの遺伝子 23,851 本[9] を使用した．リード配列は，lucy[10] を用いベクター配列除去した後の配列を利用した．表 4 のリード数に利用した配列の本数を示す．

配列のアラインメントには BLAT を用い，リード配列，近縁種遺伝子配列共に塩基をアミノ酸配列に翻訳したものを使用した．フレームは 6 フレーム全てに変換して検索している (BLAST の tblastx 相当)．アラインメントは BLAT の出すスコアで E-value が 1e-20 以下かつ，identities の値が 60 以上のものを利用した．E-value の変化をしてもかさがいの遺伝子に対応するリードの本数に変化はみられなかった．また，今回は重複遺伝子を見つけるためにミスマッチの位置を見ていかなくてはいけないので，identities のしきい値を緩くすることにした．また，アラインメントにおいて，繰り返し配列や頻繁に現れるドメイン等のため，一部の遺伝子に大量にリードが対応することがあった．このような遺伝子を防ぐため，近縁種遺伝子に対応づけられたリード数が 40 本以下の場合のみ利用している．

頻出パターン抽出の最小サポートのしきい値 c を 0.15，リード集合間の類似度を求める際のパラメータ r を 0.8，しきい値 d を 0.8 として実行した．表 4 に BLAT でカサガイの遺伝子に対応したリード数，及び，計算結果から求まった重複遺伝子数の統計を示す．いずれの種においても，カサガイに対応したリードが少なく，本実験系が進化距離としてはある程度離れた遺伝子を，近縁種として扱っている事が分かる．また，選んだ近縁種が近くないにも関わらず，オウムガイを除いて残り 3 種では 40 以上の重複遺伝子候補を見つけることができた．

3.2 考　　察

まず初めに，本手法を用いて作成されたコンセンサス配列の精度を検証する．図 5 は，近縁種遺伝子 jgi|Lotgi1|164977|fgenesh2_pg.C_sca_48000109 に対応付けられたリードのアラインメントの一部を示している．この近縁種遺伝子に対応付けられて求めた 2 本のコンセンサス配列の一部を図 6 に示す．一行目が近縁種遺伝子の配列で，二行目が zinc finger protein ではないコンセンサス配列で，三行目が zinc finger protein であるコンセンサス配列を表している．丸で囲まれた位置は変異を表していて，アラインメントの位置は図 5 と対応付けた位置である．図 5 と図 6 を比較してみると，多量の変異からも精度の高いコンセンサス配列を作成することができていることがわかる．

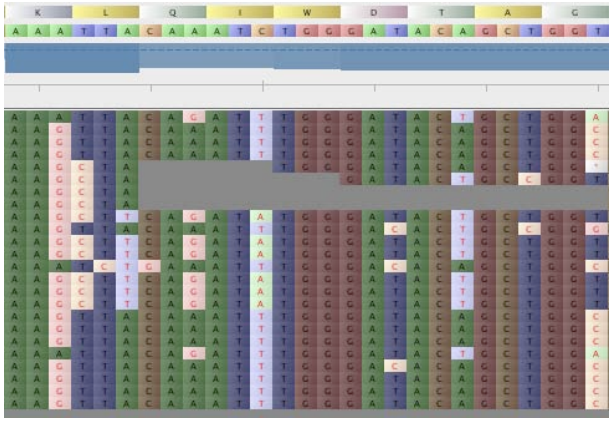

図 5　リードのアラインメント結果

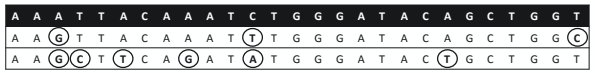

図 6　本手法で求められたコンセンサス配列

今回の解析で認められた重複遺伝子候補が，互いに異なる機能を有しているかを調査するため，個々のコンセンサス配列を NCBI の BLAST サイト[⋆1]を用いて調査した．データベースには，Non-redundant protein sequences(nr) を用い，種は指定せず，全種に対して

⋆1 http://blast.ncbi.nlm.nih.gov/Blast.cgi

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

検索を行った．また，アラインメントには BLASTX を用いた．結果の中から，ヒメイカ，ヤリイカ，ホタテガイからそれぞれ 1, 3, 1 個の重複遺伝子候補を表 5 に示す．

ヒメイカの場合, 3 本のコンセンサス配列に分かれたものが，いずれもラットでのヒストンクラスタに関連する遺伝子に対応していた．

ヤリイカで見つかった 1 つ目の遺伝子重複では，2 本のコンセンサス配列両方に共通してアクチンを制御するミオシンタンパクの重鎖に関連するが，1 本は筋肉の収縮等に関わっているのに対し，もう 1 本のコンセンサス配列では筋肉以外で関わっていることが分かっている．これは，異なる機能をもつ重複遺伝子であると予測することができる．この 2 本のコンセンサス配列は，近縁種遺伝子 jgi|Lotgi1|198383|estExt_Genewise1.C_sca_980095 に対応づけられ，当たったリード数は 11 本であった．頻出パターン解析におけるトランザクション ID の数は 188 個だったが，2.2.2 節の (a)，(b)，(c) の三つの場合における TID を削除する本手法を行うことで，64 個に削減することができた．この手法に適用した結果，複雑なアラインメント結果から重複遺伝子を予測することが可能となった．

ヤリイカで見つかった 2 つ目の遺伝子重複では，同じ種で異なる機能をもつ遺伝子重複であることが表 5 から分かる．機能の種類としては似ているが，遺伝子重複によって異なる機能をもつようになったと予測することができる．

ヤリイカで見つかった 3 つ目の遺伝子重複では, 片方のコンセンサス配列にもつ機能を，もう片方のコンセンサス配列が持っていない場合のものである．これは, 遺伝子重複によって新たな機能を獲得したものであることが推測される．

ホタテガイで見つかった遺伝子重複では, 2 本のコンセンサス配列がそれぞれ異なる機能をもっていることが分かった．

次に，5 つの場合で対応づけられた近縁種遺伝子に既に重複遺伝子群であるものがないか調査するため，JGI の検索サイト[⋆1]を用いて調査した．その結果，ヤリイカで zinc finger protein の機能の有無の違いから発見されたコンセンサス配列が対応づけられた近縁種遺伝子 jgi|Lotgi1|164977|fgenesh2_pg.C_sca_48000109 に，パラログがないことが分かった．

4. 関 連 研 究

同一種内で配列相同性の高い遺伝子はパラログと呼ばれ，データベースの整備も進むほど[11),12)] 遺伝学には重要な存在である．これらのデータベース作成手法は，ゲノムが決定さ

⋆1 http://genome.jgi-psf.org/

表 4 実行データ及び結果．アラインメントされたリード数はリードの内最低一つのカサガイの遺伝子に対応したリード数．対応遺伝子数は，最低一本のリードが対応している遺伝子数．クラスタ化したリード数は，頻出パターン解析によるリード群作成で，最低一つのリード群に含まれるリード数．重複候補の近縁種遺伝子数は，図 2 で A に相当するカサガイの遺伝子数．重複遺伝子候補数は，図 2 で A’ や A” に相当する遺伝子数．

軟体動物種	リード数	アラインメントされたリード数	対応遺伝子数	クラスタ化したリード数	重複候補の近縁種遺伝子数	重複遺伝子候補数
ヒメイカ	226,994	30,441	879	753	92	435
ヤリイカ	262,913	5,964	1,551	893	112	467
ホタテガイ	86,832	2,594	652	339	40	172
オウムガイ	232,204	1,288	419	23	4	12

表 5 重複遺伝子の詳細．5 つの遺伝子重複を挙げた．ヒメイカで 1 つ, ヤリイカで 3 つ, ホタテガイ 1 つの遺伝子重複の詳細を示している．番号はコンセンサス配列の番号を示し, そのコンセンサス配列がもつ機能と, その機能の種が示されている．Genbank ID は, その機能をもつ遺伝子の ID 番号である．

軟体動物種	番号	機能	機能の種	gene ID
ヒメイカ	1	histone cluster 2, H3c2-like	Rattus norvegicus	ref|XP_001072155.2|
	2	histone H3.3B-like	Rattus norvegicus	ref|XP_002729864.1|
	3	histone cluster 2, H3c2-like	Rattus norvegicus	ref|XP_001072155.2|
ヤリイカ	1	muscle myosin heavy chain	Loligo bleekeri	gb|ACD68201.1|
	2	non-muscle myosin II heavy chain	Loligo pealei	gb|AAK85118.1|
ヤリイカ	1	GTPase hras	Ictalurus punctatus	gb|ADO28631.1|
	2	ras-related protein rap-1b	Ictalurus punctatus	gb|ADO28745.1|
ヤリイカ	1	zinc finger protein	Ixodes scapularis	gb|EEC00705.1|
	2	unnamed protein product	Mus musculus	
ホタテガイ	1	cyclophilin-like protein	Pfiesteria piscicida	gb|ABI14283.1|
	2	peptidyl-prolyl cis-trans isomerase 5 precursor	Pediculus humanus corporis	gb|EEB15506.1|

Vol.2010-MPS-81 No.18
Vol.2010-BIO-23 No.18
2010/12/17

れ，遺伝子が分かっている種に対し，遺伝子間の相同性を調べ，相同性の高い遺伝子同士をパラログと見なす手法が取られている．しかし，これらの手法はゲノム配列全体あるいはほぼ完全な全遺伝子の配列を要求するため，決定までのハードルが高く，モデル生物に限定的な手法である．

これに対し，次世代シークエンサによる低コストな配列決定を利用し，モデル生物の近縁種に関しては，部分的な染色体重複を求める手法も利用されている[3]．しかし，この手法もモデル生物の至極近縁種に限定され，多少離れた種でも利用可能である提案手法に優位性がある．

5. まとめと今後の課題

本研究では，ゲノム未知の種から重複遺伝子を発見するために，相関ルールを用いたクラスタリング手法を導入した．本手法は，次世代シークエンサによって得られた大量配列情報と，進化的に比較的近い種のゲノム・遺伝子配列を用いる事で，変異の共起関係を発見し，それを基にリードの元となった遺伝子が同一であるか否かを判定する手法である．

本手法を次世代シークエンサの一つである Roche 社 454 で読んだ軟体動物 4 種の EST 配列に適用する事で，新たな機能を獲得した重複遺伝子を予測する事ができた．

今後は，本手法で確認できた遺伝子重複の予測が正しいかを確認するため，PCR 等を用いた周辺領域の塩基配列決定を行う予定である．

また，本手法では重複遺伝子の候補を抽出することは可能であるが，参照した近縁種の配列に重複が起こっている場合，対象種由来の重複か，近縁種由来の重複かを判別する事が難しい．近縁種の複数のゲノムの参照が可能である場合は，複数種を参照するなどの方法をとり，これらの区別を行っていくことで，予測精度を上げたいと考えている．

参 考 文 献

1) Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, Vol. 423, pp. 241, 2003.
2) Sarah A Teichmann and M Madan Babu. Gene regulatory network growth by duplication. Nature Genetics, Vol. 36, Issue 5, pp. 492, 2004.
3) Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, *et al.* Personalized copy number and segnemtal duplication maps using next-generation sequencing. Nature Genetics. Vol. 41, pp. 1061, 2009.
4) Stephen F Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. J. Mol. Biol. Vol. 215, pp. 403-10.
5) W. James Kent. BLAT — The BLAST-like alignment tool. Genome Research. Vol. 12, pp. 656-664, 2002.
6) Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. Vol. 25, Issue 14, pp. 1754-1760. 2009.
7) Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome-Biology. Vol. 10, R25. 2009.
8) Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference Santiago. Chile. 1994.
9) http://genome.jgi-psf.org/Lotgi1/Lotgi1.home.html
10) Hui-Hsien Chou and Michael H. Holmes. DNA sequence quality trimming and vector removal. Bioinformatics. Vol. 17, Issue 12, pp. 1093–1104. 2001.
11) Magalie Leveugle, Karine Prat, Nadine Perrier, Daniel Birnbaum, and Francois Coulier. ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic Acids Research. Vol. 31, No. 1. pp. 63–67. 2003.
12) Guohui Ding, Yan Sun, Hong Li, Zhen Wang, Haiwei Fan, Chuan Wang, Dan Yang, and Yixue Li. EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogon information. Nucleic Acids Research. Vol. 36, Database Issue, D255-D262. 2008.