Unique Peptide Identification of RNaseA Superfamily Sequences based on Reinforced Merging Algorithms*

TUN-WEN PAI [#,1,¶], MARGARET DAH-TSYR CHANG[#,2,¶¶], HAO-TENG CHANG[2], BO-HAN SU[1], PEI-CHIH WU[1], TAN-CHI FAN[2] and SHI-HWEI LIU[2]

[1]*Department of Computer Science, National Taiwan Ocean University, No. 2, Pei-Ning Rd. Keelung ,Taiwan 20224, Republic of China*
[¶]*twp@mail.ntou.edu.tw*

[2]*Institute of Molecular and Cellular Biology & Department of Life Science, National Tsing Hua University, No. 101, Sec. 2, Kuang Fu Rd. Hsinchu ,Taiwan 30013, Republic of China*
[¶¶]*dtchang@life.nthu.edu.tw*

Human ribonuclease A (RNaseA) superfamily consists of eight RNases with high similarity in which RNase2 and RNase3 share 76.7% identity. The evolutionary variation of RNases results in differential structures and functions of the enzymes. To distinguish the characteristics of each RNase, we developed reinforced merging algorithms (RMA) to rapidly identify the unique peptide motifs for each member of the highly conserved human RNaseA superfamily. Many motifs in RNase3 identified by RMA correlated well with the antigenic regions predicted by DNAStar. Two unique peptide motifs were experimentally confirmed to contain epitopes for monoclonal antibodies (mAbs) specifically against RNase3. Further analysis of homologous RNases in different species revealed that the unique peptide motifs were located at the correspondent positions, and one of these motifs indeed matched the epitope for a specific anti-bovine pancreatic RNaseA (bpRNaseA) antibody. Our method provides a useful tool for identification of unique peptide motifs for further experimental design. The RMA system is available and free for academic use at http://bioinfo.life.nthu.edu.tw/rma/ and http://spider.cs.ntou.edu.tw/bioinformatics/RMA.html.

Keywords: unique peptide; ribonuclease A superfamily; reinforced merging algorithm

1 Introduction

Ribonuclease (RNase) catalyzes the degradation of ribonucleic acids (RNAs). All living organisms contain RNases which mainly function during the RNA processing such as RNA maturation and turnover. Many abundant RNases such as pancreatic RNases have been isolated and well characterized. Recently, several proteins with known biological activities were found to have intrinsic ribonucleolytic activities and their amino acid sequences were homologous to that of the RNaseA superfamily.

The RNaseA superfamily is composed of proteins with similar primary structures and enzymatic activities. This class of enzymes has been suggested to participate in RNA metabolism and gene-expression regulation[1]. In humans, the genes for eight RNases including RNase1 (human pancreatic RNase)[2], RNase2 (eosinophil-derived neurotoxin)[3], RNase3 (eosinophil cationic protein)[4], RNase4[5], RNase5 (angiogenin)[6], RNase6 (RNase k6)[7], RNase7[8], and RNase8[9] are all mapped on the long arm of chromosome 14[9]. The pairwise similarities of protein sequences are more than 29%, especially RNase2 v.s.

* This work is supported by National Science Council in Taiwan, R.O.C. (NSC-93-3112-B-007-005).
# To whom correspondence. E-mail: twp@mail.ntou.edu.tw; dtchang@life.nthu.edu.tw

RNase3 (76.7%) and RNase7 v.s. RNase8 (78%). From molecular evolutionary analysis, the RNase2-RNase3 and RNase7-RNase8 gene pairs emerged as a result of a relatively recent gene duplication event[9]. Structurally, all members of the RNaseA superfamily have eight cysteines that participate in four intra-disulfide bonds, one catalytic lysine (K41) and two catalytic histidines (H12 and H119). RNase5 is an exception in that it has six cysteines instead of eight. The resolved crystal structures of RNase1, 2, 3, 4, and 5 showed similar conformations of three α-helices and five β-sheets. These features are conserved as revealed in the sequence alignments by several available programs such as BLAST and CLUSTALW. However, although the default parameters in most programs give satisfactory results, in special variables need to be taken into consideration. For example, a constrained multiple sequence alignment tool for RNaseA superfamily was developed, in order to include the structural constrains factor in multiple sequence alignments. Hence, the difference in functional properties of some RNases could be correlated[10]. To distinguish each of the human RNases in terms of sequence specificity, quantity variation, and cellular localization, monoclonal antibody (mAb) is often employed in biological investigation. Monoclonal antibody is a probe used to specifically recognize and distinguish an antigen. The recognition region, called an epitope, is often composed of 7 to 20 amino acids that are hydrophilic, charged and exposed at the surface of the protein[11-16]. However, it is still difficult to predict which region in a protein sequence is the most suitable for generation of a specific mAb, especially for those without any known structural information. In this study, we have developed reinforced merging algorithms (RMA) to rapidly identify the unique peptide motifs for each member of the highly conservative human RNaseA superfamily. Empirically, we have generated two mAbs specifically against RNase3 rather than RNase2 and RNase1. Further epitope mapping experiments revealed that the two epitopes were indeed located in the separate unique regions in RNase3 as matched by RMA.

Conventional ways to allocate the unique peptides from the RNaseA superfamily require a large number of well-defined chemical derivatives and partial segments from the protein variance for enzymatic activity tests. The complete experiments and procedures from protein sequence to monoclonal antibody and vaccine development are cost-intensive and time consuming. Therefore, it is important and useful to predict candidate peptides computationally prior to assay development. Previous methods for epitope prediction focused on distinguishing possible candidates by examining the presence of the appropriate primary/secondary anchor residues[17,18]. These prediction methods are based on the identified and published motifs for anchor identification. Lately, the prediction strategies give emphasis to pattern matching technologies, including statistical and machine learning methodologies[19,20]. However, all these developed algorithms require referencing well-known motif properties or respective structure information. Since members of the RNaseA superfamily possess strong similarities and their structural information have not been completely unveiled, we propose straightforward identification algorithms based on the proteins' local features and associated amino acid position. With the concept of amino acid substitution, continuous antigenic property enhancement, and statistical properties, different

representative levels of non-gapped unique peptide motifs were able to be extracted. More details of the proposed algorithms are introduced in the following sections.

2 Materials and Methods

The aim of this study is to create an automatic method for detecting peptides with deterministic features from a set of family protein sequences. Our main idea is to construct the deterministic unique peptide motifs by merging short primary unique patterns. A non-gapped primary unique pattern from a set of sequences is defined as a segment that possesses at least one symbol different from all possible segments in other sequences of the same length. Bottom-up merging methodologies applied on primary unique pattern strengthen the features of uniqueness from its local and neighboring characteristics. The problem definitions and system modules are introduced in the following sections.

2.1 *Problem Definition*

RNaseA superfamily sequences are represented as strings over the 20 amino acid set, and Z symbolized the set of the collected RNase sequences aforementioned in the RNaseA superfamily. The I^{th} sequence in Z will be denoted by Z_I, and the total number of such a family will be indicated by N. More specifically, the Z set is constructed as $Z = \{Z_1, Z_2,\ldots, Z_I,\ldots, Z_N\}$. Given a subsegment u in Z_I, the natural number n is defined as the length of u and is denoted by $|u|$. If u is a string and $n = |u| > 0$, $u = Z_I[i,, j]$ is the substring of Z_I by taking the i^{th} through j^{th} residues. Two different segments in a sequence Z_I can be merged if they possess overlapping residues in Z_I. Any two given segments u and v with length m and n respectively $(m,n > 0)$ can be merged with respect to Z_I and denoted by $M_{Z_I}(u,v)$. The merged result is a segment w such that $|w| = m + n - l$, where l is the number of overlapping symbols between u and v, and l is strictly greater than 0. If the number of overlapping residues is equal to **min(***m*-1, *n*-1), the merged segment w is defined as the results of strict merging operations and denoted by $SM_{Z_I}(u,v)$. The merged segment w is able to be constructed as described in Eq. (1).

$$w(i) = M_{Z_I}(u,v) = \begin{cases} u(i) & if\ \ 1 \le i \le m; \\ v(i-m+l) & if\ \ m+1 \le i \le m+n-l. \end{cases} \quad (1)$$

If u and v do not contain any overlapping residues between them, the merging operation will not change their original relation. It is clear that if u and v are merged with respect to Z_I, the length of the merged segment w is strictly less than the concatenated string uv, i.e. $|M_{Z_I}(u,v)| < |uv|$. The merging operation is proposed to enhance the discrepancy in a family set and emphasize the neighboring relationship instead of the concatenation operation. More definitions and notations are addressed as the follows,

$P_n[Z_I,.\ ]$: a set of all primary unique descriptors with segment length n in Z_I
$P_n[Z_I, j]$: the j^{th} unique descriptors with segment length n in Z_I
(*i.e.* $P_n[Z_I,.\ ] = \bigcup_j P_n[Z_I, j]$)

$M(P_n[Z_I, \cdot])$: a set of merged unique descriptors with segment length n over Z_I.

$P_n^m[Z_I, j]$: a set of all primary patterns with reserving m symbols from the $P_n[Z_I, j]$ unique descriptor

$P_{n,k}^m[Z_I, j]$: the k^{th} pattern in set $P_n^m[Z_I, j]$

$C(n, m)$: the number of possible combinatorial pattern set of reserved m symbols from a primary unique pattern with length n

$R(P_n^m[Z_I, j])$: the representative percentage of the j^{th} primary unique descriptor $P_n[Z_I, j]$ with m symbols retained in family set Z and shown in Eq. (2)

$$R\left(P_n^m[Z_I, j]\right) = 1 - \frac{1}{C(n,m)} \sum_{k=1}^{C(n,m)} \frac{\left\langle P_{n,k}^m \right\rangle - 1}{N-1} \tag{2}$$

where $\left\langle P_{n,k}^m \right\rangle$ is the number of the appearing count of the k^{th} pattern $P_{n,k}^m[Z_I, j]$ shown in the set Z and $\left\langle P_{n,k}^m \right\rangle \in [1,N]$. The number of $\left\langle P_{n,k}^m \right\rangle$ can be calculated by counting the appeared condition of the k$^{\text{th}}$ pattern $P_{n,k}^m[Z_I, j]$ in each Z_I, $I \in [1,N]$. The value of $\left\langle P_{n,k}^m \right\rangle$ is greater or equal to 1, since there must be at least one sequence Z_I from set Z which contains the complete unique subsegment $P_n[Z_I, j]$. If the k^{th} tolerant pattern $P_{n,k}^m[Z_I, j]$ can be matched in each sequence in Z, then $\left\langle P_{n,k}^m \right\rangle = N$ and $\frac{\left\langle P_{n,k}^m \right\rangle - 1}{N-1} = 1$. Furthermore, if different patterns can be found for the k in every sequence as well, then $\frac{1}{C(n,m)} \sum_{k=1}^{C(n,m)} \frac{\left\langle P_{n,k}^m \right\rangle - 1}{N-1} = 1$ and $R\left(P_n^m[Z_I, j]\right) = 1 - 1 = 0\%$. This describes the special condition that all possible tolerant patterns of unique descriptor $P_n^m[Z_I, j]$ are able to be discovered in each sequence of Z. Therefore, its unique representative percentage will be decreased to 0%. On the other hand, if all the combinatorial tolerant patterns from $P_n^m[Z_I, j]$ cannot be found in other family sequences, then $\left\langle P_{n,k}^m \right\rangle = 1$ for k = [1,$C(n,m)$] and $\frac{1}{C(n,m)} \sum_{k=1}^{C(n,m)} \frac{\left\langle P_{n,k}^m \right\rangle - 1}{N-1} = 0$, and it follows that $R\left(P_n^m[Z_i, j]\right) = 1 - 0 = 100\%$. This describes another special case that provides all of the possible tolerant patterns for distinguishing in which subsegment of $P_n^m[Z_I, j]$ still possesses 100% unique representative percentage. These subsegments will be in a sense of optimal condition for representing a sequence from a family set. Without loss of generality, the ratio of $R\left(P_n^n[Z_I, j]\right)$ is equal to 100%, since $P_n^n[Z_I, j]$ is initially defined as a unique descriptor from the family set Z.

2.2 *System Configuration and Description*

Three main phases are designed in the system including grouping, searching, and merging modules. The main purpose of the grouping module is to discriminate the tolerant feature from unique peptide motif representation. Since the possible substitutions of different amino acids from observed frequencies in algorithms of related proteins are well-defined[21,22], we provide a fundamental clustering algorithm to assemble 20 amino acids into

several independent groups. The clustering algorithms are based on hierarchical methodology (agglomerative algorithms) which produces a series of decreasingly clustered numbers at each step[23]. The algorithms merge two clusters from the previous step into one and stop merging when the threshold value setting is met. For example, if users select the BLOSUM62 substituting matrix and set a thresholding value of 3 for clustering, the module will cluster "I(Isoleucine)" and "V(Valine)" into the same group, and "F(Phenylalanine)" and "Y(Tyrosine)" into another group. All other amino acids represent one individual group. Since (I,V) and (F,Y) possess a higher substitution possibility than the others, the grouped amino acids in following comparisons of unique peptide motifs will be considered as possessing identical characteristics.

For the second searching module, the system requires the length of the primary segment and the tolerance number in mismatching situation as parameters. Based on these settings, the module performs modified Boyer Moore algorithms[24] to filter out all possible primary unique patterns, and their representative percentages are calculated. Primary unique pattern is defined as the basic elements for the next merging operations. Since each set of grouped amino acids is represented by a unique symbol, the algorithm examines all candidate patterns and extracts the positional information of primary unique patterns by scanning all substituted sequences. If the clustering results are less than 20 groups from the previous module, the tolerant feature of uniqueness will be discriminated and less primary unique patterns can be extracted from the family set. For example, if I and V are clustered in a group according to BLOSUM62 substitution matrix, then pattern IRF and VRF will be considered as resembled patterns and do not possess strict unique properties. Thus, neither one is selected as primary unique pattern. In addition to adopting clustered results to describe the features of uniqueness, this module also provides another statistical analysis to show the level of determined characteristics for each extracted unique pattern. Users are able to assign the mismatched number, which is less than the length of a primary unique pattern. If the length of the primary unique pattern is n and reserved m symbols for each pattern, i.e. allow (n-m) mismatches from matching processes, then the representative level of uniqueness can be represented by $R\ (P_n^m[Z_1, j])$ defined in Sec. 2.1. All searched primary unique patterns with different representative percentages are then examined by the next module to perform merge operations based on their neighboring conditions and representative levels of uniqueness.

In the last merging phase, the system provides reinforced merging processes to obtain predicted unique peptide motifs. Users are able to specify the minimal length of final merged segments with local primary segments that satisfy the assigned representative percentage. The RMA system provides four different merging methods to enhance different uniqueness levels including MERGE (M), TRIM MERGE (TM), STRICT MERGE (SM), and TRIM STRICT MERGE (TSM) operators. The MERGE operator, $M(u,v)$, provides the results of unique peptide motif searches with a degree value above the lowest limit of primary pattern length, representative percentage, and overlapping criteria. The TRIM MERGE operator, $TM(u,v)$, performs MERGE operations and TRIM processes. The trimming function returns a substring from each merged segment with the symbols stripped off the beginning and the end. The boundary residues will be trimmed off when they are matched with another merged unique peptide. For example, both CHHSGSQVP and CHRSRFRVP are merged unique patterns from two different protein sequences with

primary unique pattern length of 3. Therefore, the system will evaluate 2 residues at both the beginning and end positions of all merged segments. In this case, the final unique patterns become HSGSQ and RSRFR after trimming processes. The STRICT MERGE operator, $SM(u,v)$, performs strict merging operations to ensure each subsegment of merged peptides satisfy the lower limit of uniqueness level. The last TRIM STRICT MERGE operator, $TSM(u,v)$, performs trimming functions after the strict merge operator and provides the most rigorous properties of uniqueness such that all possible partial segments of the merged substring will satisfy the representative percentage requirement as well.

Once the merged unique peptide motifs are determined, each unique peptide motif is ranked according to priority weighting measurements. Four different features of support are considered in this process, and each characteristic is essential to assist in ranking the searched unique peptide motifs in favor of serving as epitopes for generation of antibodies. The first important property is the hydrophilicity[11] of the residues, and a weighting coefficient of 4 is given to eleven such residues. The second property is the charge[12,13] of the residues, and a weighting coefficient of 3 is assigned to three positively and two negatively charged amino acids. The third one is the number of secondary amines[14,15] in the motifs and, hence, a weighing coefficient of 2 is given to the Proline residue. The last property is the proximity of the motifs towards the N or C terminal[16] of the protein for which a weighting coefficient of 1 is assigned for any residue located within 20 amino acids from either end of a protein. Based on the criteria, relative weighting measurement of each amino acid can be calculated by verifying the four properties with the sum of its coefficients. It is obvious that the highest weighting coefficient will be 8 and the lowest value will be 0 for any amino acid. Table 1 provides the weighting coefficients for each amino acid according to these four properties. In order to distinguish the priority of searched unique peptide motifs, the final scores for merged segments are calculated, normalized for ranking process, and evaluated by the calculation of Eq. (3).

$$S(w) = \frac{\sum_{i=1}^{|w|} SA(w(i))}{|w|} \tag{3}$$

where $S(w)$ is the normalized score for unique peptide motif w, $SA(w(i))$ is the weighting score for the i^{th} amino acid $w(i)$, and $|w|$ is the length of w. Each unique peptide motif will be normalized in the range [0,8] for ordering. The higher scores in the rank indicate better antigenicity of the unique peptide motif; which suggest that the motif may serve as a potential epitope for generation of antibodies.

Table 1. The weighting coefficients for 20 amino acids respectively.

Amino Acids	Hydrophilic	Charged	2° Amine	N or C Terminus[a]	Weighting Scores
Alanine (A)	0	0	0	0/1	0/1
Arginine (R)	4	3	0	0/1	7/8
Asparagine (N)	4	0	0	0/1	4/5
Aspartic acid (D)	4	3	0	0/1	7/8
Cysteine (C)	4	0	0	0/1	4/5
Glutamine (E)	4	0	0	0/1	4/5
Glutamic acid (Q)	4	3	0	0/1	7/8
Glycine (G)	4	0	0	0/1	4/5
Histidine (H)	4	3	0	0/1	7/8
Isoleucine (I)	0	0	0	0/1	0/1
Leucine (L)	0	0	0	0/1	0/1
Lysine (K)	4	3	0	0/1	7/8
Methionine (M)	0	0	0	0/1	0/1
Phenylalanine (F)	0	0	0	0/1	0/1
Proline (P)	0	3	2	0/1	5/6
Serine (S)	4	0	0	0/1	4/5
Threonine (T)	4	0	0	0/1	4/5
Tryptophan (W)	0	0	0	0/1	0/1
Tyrosine (Y)	4	0	0	0/1	4/5
Valine (V)	0	0	0	0/1	0/1

[a] Amino acids in searched unique peptide motifs located at either N or C terminus is given a weighting coefficient of 1.

2.3 *Analysis of protein sequences by RMA*

The human RNase1, 2, 3, 4, 5, 6, 7 and 8 and bpRNaseA were saved in FASTA format and entered into the RMA system for analysis using manually set parameters. The parameters for length of lower bound, upper bound and reserved sites were set as 3, 3, and 3 for demonstration. After employing merge and strict merge processes, the merged unique peptide motifs with length greater than 15 and 6 were identified and labeled in red and blue series, respectively. The conserved regions were labeled in black in both cases.

2.4 *Molecular Cloning*

The DNA fragments, mature RNase2, mature RNase3, RNase3$_{1\text{-}73}$, RNase3$_{24\text{-}133}$, and RNase3$_{51\text{-}73}$, were prepared by PCR and subsequently cloned into plasmids. The mature RNase2 and RNase3 inserts were ligated into the pRSETA and pET3a vectors, respectively. The other gene fragments were fused to GFP in pET23a (+) vectors. Between RNase3 fragments (RNase3$_{1\text{-}73}$ and RNase3$_{51\text{-}73}$) and GFP, a linker (222 bps) was inserted. The recombinants pET3a-mRNase3-H$_6$, pRSETA-mRNase2, pET23a-RNase3-GFP-H$_6$, pET23a-RNase3$_{1\text{-}73}$-GFP-H$_6$, pET23a-RNase3$_{24\text{-}133}$-GFP-H$_6$, pET23a-RNase3$_{51\text{-}73}$-GFP-H$_6$ and pET23a-RNase3$_{24\text{-}50}$-GFP-H$_6$ were transformed into *E. coli* BL21 (DE3) for protein expression.

2.5 *Protein Purification, Monoclonal Antibody Generation and Western Blotting*

The inclusion bodies of recombinant proteins were isolated and resuspended in binding buffer under denaturing conditions. After centrifugation, the supernatant was loaded into

His-Bind® metal chelating chromatography (Novagen). To remove other proteins, the column was washed with binding buffer and wash buffer. The recombinant RNase3 was eluted with elution buffer, and the purified proteins were analyzed by 15% SDS-PAGE stained with Coomassie Blue. BALB/c mice were immunized with 10 μg recombinant mature RNase3 suspended in Freund's complete adjuvant. These animals were boosted twice with the same amount of the antigen in Freund's incomplete adjuvant after a 2-week interval. The final boost was performed by injection with 10 μg of the antigen without adjuvant. Three days after the last immunization, splenocytes were collected and fused with NS-1 myeloma cells. Hybridoma producing anti-RNase3 mAb were obtained by indirect ELISA screening with purified recombinant RNase3 as the antigen while recombinant RNase2 was used as the negative control. Positive clones were subjected to single-cell cloning twice by the limiting dilution method. The mAbs were purified from hybridoma supernatant using the Montage® Antibody Purification Kit (Millipore) and named as mAb 109 and mAb 2450. For detection of proteins, the proteins were transferred to a nitrocellulose membrane which was incubated in 3% non-fat milk at room temperature for blocking. One hour later, the Ab was diluted with 1.5% non-fat milk and reacted with the target protein at room temperature for 2 hr. The membrane was washed with TTBS (20 mM Tris-HCl, pH 7.6, 0.8% NaCl, 1% Tween-20) three times for 15 min each. The secondary Ab was then diluted in the 1.5% non-fat milk and reacted with the first Ab for 2 hr. The membrane washing procedure was repeated three times, and the target proteins were visualized using Super Signal West Pico (Pierce).

2.6 *Parameter setting for DNAStar*

The amino acid sequences of query proteins were saved in PROTEIN format and executed by DNAStar. The sequential sequences greater than 0.5 antigenic index were selected and compared with the unique peptide motifs identified by RMA. The coverage of antigenic regions and unique peptide motifs was selected to be more than 50%.

3. Results

3.1. *Specificity of Anti-RNase3 Monoclonal Antibodies*

The recombinant constructs were transformed into *E. coli* BL21 (DE3) and expressed by induction with 0.5 mM IPTG. The recombinant RNase3 was purified with metal chelating chromatography under denaturing conditions (Fig. 1A). The purity of recombinant RNase3 was more than 95% and sufficient for mAb production. The monoclonal antibodies against RNase3 were successfully produced. To analyze the specificity of anti-RNase3 mAbs, the RNase1, BSA and bacterial extract containing recombinant RNase3-H_6, RNase3-GFP-H_6, GFP-H_6, and RNase2 were separated by 15% SDS-PAGE (Fig. 1B, upper panel). The proteins were transferred onto a nitrocellulose membrane and probed by mAb 109 and mAb 2450, separately (Fig. 1B, middle and lower panels). Fig. 1B indicated that these two mAbs could specifically recognize RNase3-H_6 and RNase3-GFP-H_6, but did not cross-reacted to

RNase1, RNase2 or GFP tag. Both mAbs 109 and 2450 could specifically distinguish RNase3 from RNase2, the two highly homologous human RNases.

3.2. *Epitope Mapping of Anti-RNase3 Monoclonal Antibodies*

For efficient expression, the mature or truncated RNase3 was fused to a GFP-H_6 tag and named as RNase3-GFP-H_6, RNase3_{1-73}-GFP-H_6, RNase3_{24-133}-GFP-H_6, and RNase3_{51-73}-GFP-H_6 (Fig 2A). These proteins could be expressed in *E. coli* and recognized by anti-His antibody as a positive control (Fig. 2B, upper left panel). After probing with mAb 109 revealed four single-band signals, indicating that the epitope of mAb 109 was located within RNase3_{51-73} (Fig. 2B, middle panel). By the same method, mAb 2450 could probe RNase3-GFP-H_6, RNase3_{1-73}-GFP-H_6, and RNase3_{24-133}-GFP-H_6, but not RNase3_{51-73}-GFP-H_6 (Fig 2B, lower left panel). Based on the experimental results, the epitope of mAb 2450 was suggested to be located within RNase3_{24-50}. In order to confirm the epitope of mAb 2450, the bacterial lysate containing recombinant RNase3_{24-50}-GFP-H_6 was probed by anti-His antibody and mAb 2450, separately. Fig. 2B showed that the mAb 2450 indeed recognized the RNase3_{24-50} fragment. According to the previously mentioned experiments, both mAbs could specifically distinguish RNase3 from RNase2, and the two mAbs recognized different epitopes, indicating at least two distinguishable regions existed between RNase2 and RNase3.

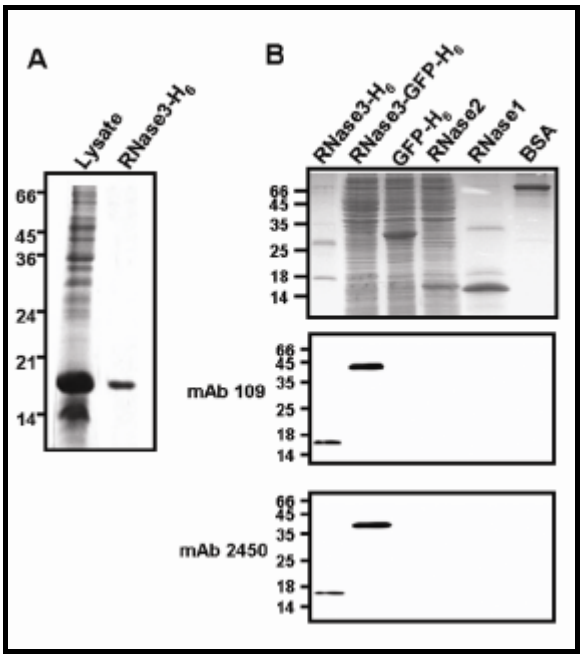


Fig. 1. The purification of mature RNase3 and specificities of anti-RNase3 monoclonal antibodies. The mature RNase3 was purified with metal chelating chromatography and located at 18 kDa (A). The RNase3-H_6, RNase3-GFP-H_6, GFP-H_6, RNase2, RNase1 and BSA were analyzed by 15% SDS-PAGE and stained with Coomassie blue (B, upper panel). Both mAb 109 and 2450 could probe RNase3-H_6 and RNase3-GFP-H_6, but not GFP-H_6, RNase2, RNase1 and BSA (negative control) (B, middle and lower panel).

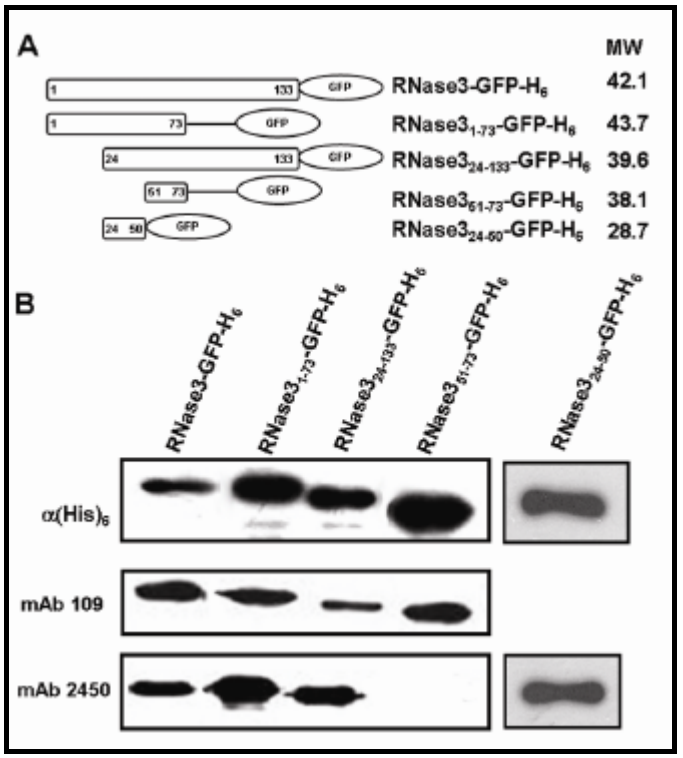


Fig. 2. Epitope screening of anti-RNase3 monoclonal antibodies. The cartoon of both mature and truncated RNase3 fused to GFP-H_6 was shown. The molecular weight was calculated with Compute pI/Mw tool (A). Anti-His antibody could probe all the recombinant proteins to prove the protein expression and loading control (B, upper panel). The mAb 109 could recognize the first four recombinant proteins (B, middle panel) and the mAb 2450 could probe the first three proteins and $RNase3_{24-50}$-GFP, but not $RNase3_{51-73}$-GFP-H_6 (B, lower panel).

3.3. *The Output of Protein Sequences Analyzed by RMA*

RMA program was applied to the protein sequences of eight members of human RNaseA superfamily. In Fig. 3, the unique peptide motifs with length greater than 6 were labeled in blue and light blue, and the black characters represented the conserved regions in the eight RNases. The results revealed that these color coded regions represented the non-gapped sequential unique peptide motifs possibly the most distinguishable areas among the members of human RNaseA superfamily. The key residues in the ribonucleolytic active sites were located in the conserved regions among most of the RNases (Fig. 3, boxed). The epitopes of mAb 109, mAb 2450 and Boix's Ab[21] were underlined in green, red, and yellow, respectively. The unique peptide motif, DPRDSPRY, identified by RMA matched well with the epitope designed for Boix's Ab. It was thus expected that other unique peptide motifs, TIAMRA, NYRWRC, and SIRCPHNRTLNN were probably recognized by different antibodies.



Fig. 3 Sequence comparison of eight human RNases by strict merge operators with length of unique peptide motifs greater than 6 as shown in blue and light blue. The three conserved active sites for ribonucleolytic activity are boxed. The red, green and yellow underlines in RNase3 represent the epitopes of mAb 2450, mAb 109 and Boix's Ab, respectively. The corresponding regions to the epitopes of three antibodies are dashed-underlined in the other RNases.

3.4. *Application of RMA on homologous proteins from different species*

The sequence comparison of homologous proteins among different species provides information on molecular and functional evolution. In this study, hpRNase and its bovine homologue bpRNaseA were analyzed by RMA. The result shown in Fig. 4 revealed that 6 unique peptide motifs of 10 to 12 residues in length were identified in each sequence. We found that a unique peptide motif M29-C40 in bpRNaseA was included in the recognition site (S32-V43) of a polyclonal Ab αbpRNaseA generated by Younus and colleagues[25]. Our results demonstrated that RMA could be applied to identify the unique peptide motifs in highly homologous protein families fromsame or different species.

[bpRNaseA.txt]
MALKSLVLLSLLVLVLLLVRVQPSLGKETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKN
GQTNCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV αbpRNaseA
[hpRNase1.txt]
MALEKSLVRLLLVLILLVLGWVQPSLGKESRAKKFQRQHMDSDSSPSSSSTYCNQMMRRRNMTQGRCKPVNTFVHEPLVDVQNVCFQEKVT
CKNGQGNCYKSNSSMHITDCRLTNGSRYPNCAYRTSPKERHIIVACEGSPYVPVHFDASVEDST

Fig. 4. Sequence comparison of homologous RNaseAs in bovine and human. The unique peptide motifs are labeled in blue and the epitope for αbpRNaseA is underlined.

3.5. *The Comparison of Unique Peptides Identified by RMA and Epitopes Predicted by DNAStar*

As shown in Fig. 3, sixty-seven unique peptide motifs were identified by RMA. In this study, we have further compared the unique peptide motifs identified by RMA with the potential epitope regions predicted by DNAStar[26], which can predict the potential antigenic/epitope regions in one input protein sequence. Within the 12, 7, 8, 12, 11, 8, 6, and 3 unique peptide motifs in RNase1, 2, 3, 4, 5, 6, 7, and 8 identified by RMA, 6, 2, 6, 7, 8, 2, 4 and 2 motifs correlated well with the potential epitopes predicted by DNAStar, indicating that more than 55% (37/67) of unique peptide motifs could be predicted as potential epitopes respectively. Empirically, we have identified the epitopes of mAb 109 and mAb 2450 to be located in V51-R73 and T24-N50 in RNase3, respectively. The epitopes of mAb 109, mAb 2450 and a polyclonal Ab generated by Dr. Boix's[27] have been found to match perfectly with the potential epitopes predicted by DNAStar. Table 2 revealed that within the eight unique peptide motifs with higher ranking identified by RMA, six of them correlated well with the potential epitopes predicted by DNAStar. Therefore, the RMA could identify the unique peptide motifs in each member of a protein family, and the DNAStar could predict the potential epitopes in one member of a protein family. Sequence analyses employing the combination of these two methods may provide additional valuable information. Table 2 suggested that the unique peptide motif Y_{98}ADRPGRRFYV$_{108}$ identified by RMA and the epitope R_{97}YADRPG$_{103}$ predicted by DNAStar may serve as a peptide antigen to produce antibody which recognizes and distinguishes RNase3 from other RNases in the superfamily. Our results also indicated that the correspondent motifs may be applied to generate specific antibodies against an individual human RNase.

Table 2. Comparison of unique peptides identified by RMA and epitopes predicted by DNAStar in RNase3.

Unique peptides (ordered by ranking) [a]	Predicted epitopes [a]	Experimental epitopes
D_{115}PRDSPRY$_{122}$	A_{110}CDNRDPRDAPRY$_{122}$	Boix's Ab[b]
N_{32}YRWRC$_{38}$	W_{35}RCKNQ$_{40}$	mAb 2450[c]
H_{72}RSRFRVP$_{79}$	N_{69}NCHRSRF$_{76}$	mAb 109[d]
S_{59}IRCPHNRTLNN$_{70}$	I_{60}RCPHNRT$_{67}$	mAb 109[d]
Y_{98}ADRPGRRFYV$_{108}$	R_{97}YADRPG$_{103}$	
G_{-11}LMGVE$_{-6}$[e]		
D_{84}LINPGAQN$_{92}$		
H_{15}ISLNPPRCTIAMRA$_{29}$	N_{19}PPRC$_{23}$	

[a] The assigned numbers start from the first residue, Arg, of mature RNase3.
[b] D115-Y122 determined by Boix.
[c] T24-N50 experimentally determined by this work.
[d] V51-R73 experimentally determined by this work.
[e] Signal peptide upstream of mature RNase3

4. Discussion

The determination of epitopes in a protein is important for generation of peptide antigens or antibody drugs. The epitopes should possess characteristic properties such as hydrophilicity, antigenicity, and specificity. Most available programs predict the epitopes according to such properties. In our RMA program, we discriminated the accumulation of unique

features to distinguish different regions from eight RNases. The unique sequential sequences labeled by blue and dark blue in Fig. 3 distinguished RNase3 from all other RNases. Two mAbs against RNase3 were successfully generated and characterized to specifically recognize RNase3 but not the highly conservative RNase2. Furthermore, the three-dimensional structure of RNase3 revealed that three epitopes were all exposed at the surface, and the correspondent regions in RNase1, RNase2, and RNase4 labeled in red, green, and yellow were evidently different in their structures (Fig. 5). The epitope of mAb 2450 was determined to be located in the T24-N50. These 27 amino acids contained two unique peptide motifs, T24-A29 and N32-C38. The latter, but not the former was considered to be the potential epitope as predicted by DNAStar. RMA has also been demonstrated to be useful for analyzing homologous proteins from different species in addition to different proteins of one family in a particular species. The analysis among different species may reveal the residue variation during molecular evolution. To our knowledge, sometimes an antibody may not be successfully produced by injecting the immunized animals with a designed peptide or protein, probably due to the lack of the uniqueness in the antigenic sequence[28]. Our RMA could identify the unique peptide motifs which do not exist in the homologous protein in the immunized animals. This method provides an initial screening for potential epitope design to raise the possibility of obtaining polyclonal antibodies against synthetic peptides. Hence RMA can be used as a novel tool to identify unique peptide(s) and potential antigenic regions among several highly conserved proteins superfamilies.

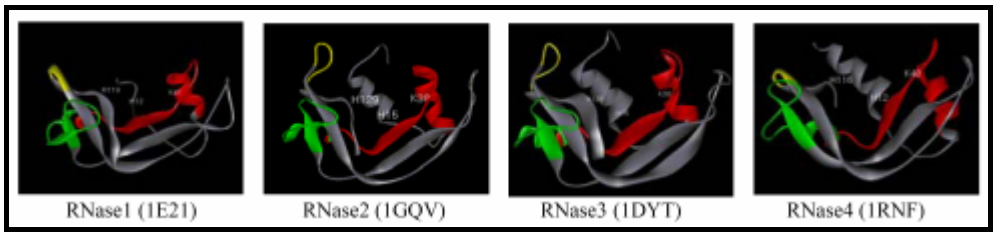


Figure 5. The three-dimensional structures of human RNase1 (N7-truncated), RNase2, RNase3 and RNase4. The structure of RNase A superfamily contains 3 α-helices and 5 β-sheets. The catalytic residues of RNase, His, Lys and His (white) are labeled. The epitope regions of mAb 109, mAb 2450 and Boix's Ab are labeled in each structure at correspondent positions in green, red and yellow, respectively.

Acknowledgements

The authors thank Drs. P.-C. Huang (Department of Biochemistry and Molecular Biology, the Johns Hopkins University), Y.-K. Lai, L.-Y. Lin, Y.-J. Sun, (Department of Life Science, NTHU), C.-Y. Tang (Department of Information Science, NTHU), and Miss I.-P. Lin and C. Teng for critical comments and proofreading.

5. References

1. G. D'Alessio, "New and cryptic biological messages from RNases," *Trends Cell Biol.* **3**(4), 106-109 (1993).
2. J. L. Weickmann, M. Elson and D. G. Glitz, "Purification and characterization of human pancreatic ribonuclease," *Biochemistry* **20**(5), 1272-1278 (1981).

3. N. R. Slifman *et al.*, "Ribonuclase activity associated with human eosinophil-derived neurotoxin and eosinophil cationic protein," *J. Immunol.* **137**, 2913-2917 (1986).
4. I. Olssom and P. Venge, "Cationic proteins of human granulocytes. I. Isolation of the cationic proteins from the granules of leukaemic myeloid cells," *Scand. J. Haematol.* **9**, 204-214 (1972).
5. H. F. Rosenberg and K. D. Dyer, "Human ribonuclease 4 (RNase 4): coding sequence, chromosomal localization and identification of two distinct transcripts in human somatic tissues," *Nucleic Acids Res*. **23**(21), 4290-4295 (1995).
6. D. J. Strydom *et al.*, "Amino acid sequence of human tumor derived angiogenin," *Biochemistry* **24**(20), 5486-5494 (1985).
7. H. F. Rosenberg and K. D. Dyer, "Molecular cloning and characterization of a noval human ribonuclease (RNase k6): increasing diversity in the enlarging ribonuclease gene family," *Nucleic Acids Res.* **24**, 3507-3513 (1996).
8. J. Zhang, K. D. Dyer and H. F. Rosenberg, "Human RNase 7: a new cationic ribonuclease of the RNase A superfamily," *Nucleic Acids Res*. **31**(2), 602-607 (2003).
9. J. Zhang, K. D. Dyer and H. F. Rosenberg, "RNase 8, a novel RNase A superfamily ribonuclease expressed uniquely in placenta," *Nucleic Acids Res*. **30**(5), 1169-1175 (2002).
10. C. Y. Tang, C. L. Lu, M. D. T. Chang, Y. T. Tsai, Y. J. Sun, K. M. Chao, J. M. Chang, Y. H. Chiou, C. M. Wu, H. T. Chang and W. I. Chou, "Constrained multiple sequence alignment tool development and its application to RNase family alignment," *J. Bioinform. Comput. Biol.* **1**(2), 267-287 (2003).
11. G. W. Welling, W. J. Weijer, van der Zee R and S. Welling-Wester," Prediction of sequential antigenic regions in proteins," *FEBS Letters* **188**(2), 215-218 (1985).
12. W. A. Stylos, C. F. Merryman and P. H. Maurer, "Antigenicity of polypeptides (poly-alpha-amino acids). Distribution of sheep antibodies to polymers of alpha-L-amino acids of varying net electrical charge and lysozyme," *Int. Arch. Allergy Appl. Immunol*. **39**(4), 381-390 (1970).
13. P. H. Mauere, C. F. Merryman and W. A. Stylos, "Antigenicity of polypeptides (poly-alpha-amino acids). Distribution of sheep antibodies to dinitrinophenyl-(DNP) conjugates of polymers of alpha-L-amino acids of varying electrical charge," *Int. Arch. Allergy Appl. Immunol*. **39**(4), 435-444 (1970).
14. H. E. Jasin and L. E. Glynn, "The antigenticity of some synthetic Poly-Iminoacids. I. Antigenic properties of Poly-L-Proline," *Immunology* **42**, 95-105 (1965).
15. H. E. Jasin and L. E. Glynn, "The antigentic properties of some synthetic Poly-Iminoacids. II. The antigenicity of polypeptides related to collagen; peptides containing hydorxyproline and acetyl-hydroxy-proline," *Immunology* **32**, 260-269 (1965).
16. T. H. Watts, P. A. Sastry, R.S. Hodges and W. Paranchych, "Mapping of the antigenic determinants of Pseudomonas aeruginosa PAK polar pili," *Infect. Immun.* **42**(1), 113-121 (1983).
17. E. J. Collins, D. N. Garboczi and D. C. Wiley Collins, Three-dimensional structure of a peptide extending from one end of a class I MHC binding site," *Nature* **371**(6498), 626-629 (1994).
18. J. Ruppert *et al.*, "Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules," *Cell* **74**(5), 929-937 (1993)
19. P. Donnes and A. Elofsson, "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics* **3**(1), 25 (2002).
20. K. Gulukota *et al.*, "Two complementary methods for predicting peptides binding major histocompatibility complex molecules," *J. Mol. Bio.l* **267**(5), 1258-1267 (1997).
21. M. O. Dayhoff, R. M. Schwartz and B.C. Orcutt, "A model of evolutionary change in proteins. matrices for detecting distant relationships," National biomedical research foundation, Washington DC, pp. 345-358
22. S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks. *Proc Natl. Acad. Sci. USA* **89,** 10915-10919 (1992).
23. E. M. Voorhees, "Implementing agglometrative hierarchical clustering algorithms for use in document retrieval, " Information Processing & Management, 22, 465-476 (1986).

24. R. S. Boyer and J. S. Moore, "A fast string searching algorithm," *Commun. ACM* **20,** 762-772 (1977).
25. H. Younus, M. Owais, D. N. Rao and M. Saleemuddin, "Stabilization of pancreatic ribonuclease A by immobilization on Sepharose-linked antibodies that recognize the labile region of the enzyme," *Biochim. Biophys. Acta,* **1548**(1), 114-120 (2001).
26. T. N. Plasterer, "PROTEAN. Protein sequence analysis and prediction," *Mol. Biotech.* 16, 117-126 (2000).
27. E. Boix *et al.*, "Identification and characterization of human eosinophil cationic protein by an epitope-specific antibody," *J. Leukoc. Biol.* **69**(6), 1027-1035 (2001).
28. E. Harlow and D. Lane, *Antibodies: a laboratory manual,* Cold Spring Harbor Laboratory, New York, NY, (1988).



TUN-WEN PAI received the degree in Electrical Engineering from National Taipei University of Technology, Taipei, Taiwan, in 1984, the M.S.E. degree in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, MD, in 1989, and the Ph.D. degree in Electrical and Computer Engineering from Duke University, Durham, NC, in 1993. He is currently an associate Professor of Computer Science Dept. at National Taiwan Ocean University, Keelung, Taiwan. His research interests are in algorithms and systems for multimedia, modeling and synthesis of cartoon caricatures, and biological sequence analysis.



MARGARET DAH-TSYR CHANG received the B.S. degree in Chemistry from National Taiwan University at Taipei, Taiwan in 1987 and the PhD degree in Chemistry from the Johns Hopkins University at Baltimore, Maryland, USA in 1993. She is currently an associate professor in the Institute of Molecular and Cellular Biology & Department of Life Science, National Tsing Hua University at Hsinchu, Taiwan. Her current research interests include investigation on molecular and cellular mechanisms of human proteins related to diseases, characterization of structure-function relationship of recombinant proteins, and integration of protein chemistry and bioinformatics.



HAO-TENG CHANG is currently a PhD student of Department of Life Science in National Tsing Hua University, Hsinchu, Taiwan. He received his B.S. degree in Biology from Chinese Culture University at Taipei, in 1998 and the M.S. degree in Life Science from NTHU in 2000. His current research focuses on the protein-protein interaction and cell entry of RNase.



BO-HAN SU is currently a master student of Computer Science Department, National Taiwan Ocean University. He received his B.E. in Computer Science from the National Taiwan Ocean University in 2004. His current research focuses on the computational algorithms to predict significant features from protein 3D structures and to identify the group features for multiple biological sequences.



PEI-CHIH WU is currently a master student of Department of Computer Science at National Twain Ocean University. She received her B.E. degree in Computer Science from National Twain Ocean University in 2003. Her current research interests cover the analysis of biological sequences to identify the unique peptide segment motifs and development of approximate matching algorithms.



TAN-CHI FAN is currently a Ph.D. student of Life Science, National Tsing Hua University. She received her B.S. in Department of Life Science from National Tsing Hua University in 1999. Her current research focuses on the mechanism of ribonuclease trafficking and the role this process which plays in regulating signal transduction.



SHI-HWEI LIU is now a Ph.D. student in the Department of Life Science, National Tsing Hua University, Taiwan. His current research interests include genetic engineering, molecular and cell biology.