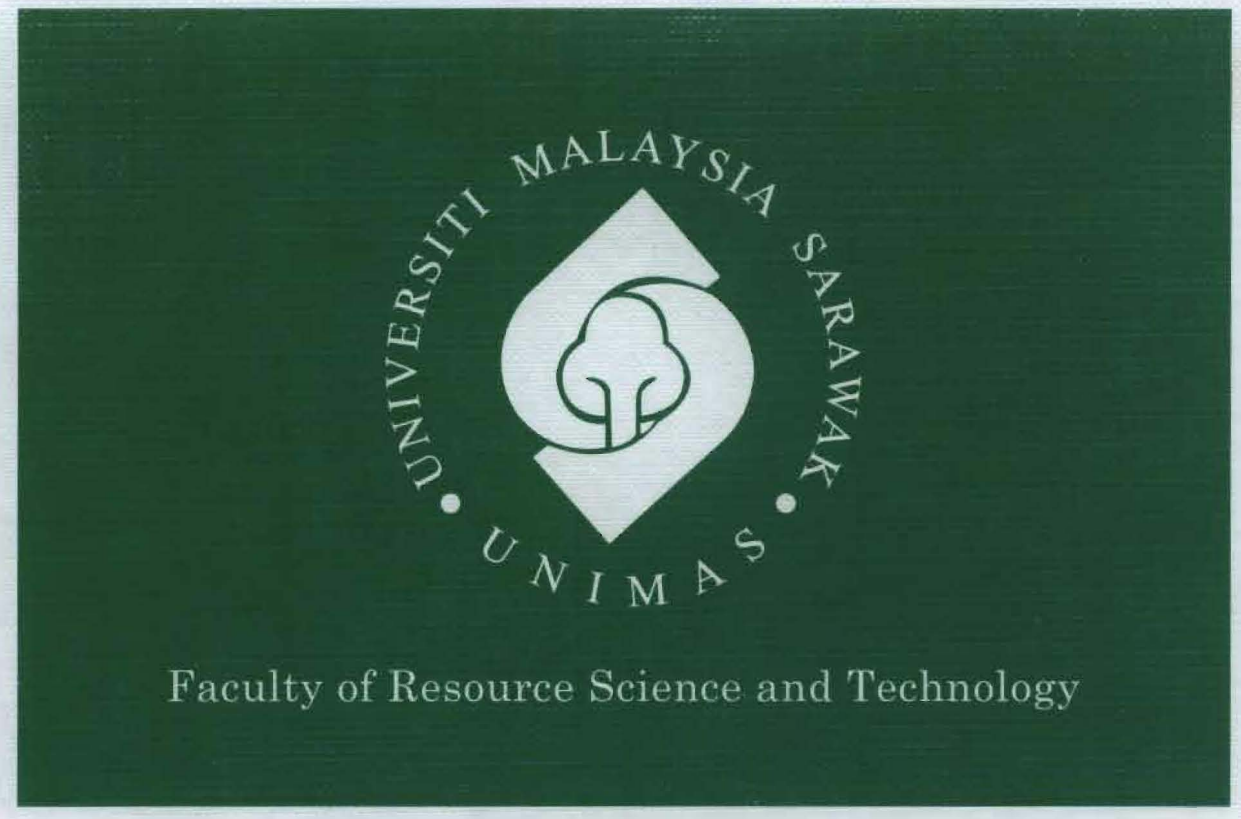

BIOINFORMATICS ANALYSIS OF THE RIBOSOMAL PROTEIN, RPL27, RPL37a AND RPL41: 3-D PROTEIN MODELING AND PROTEIN-PROTEIN INTERACTION PREDICTION

Stella Chan Li Li

Bachelor of Science with Honours
(Resource Biotechnology)
2012

Bioinformatics Analysis of the Ribosomal Proteins, RPL27, RPL37a and RPL41: 3-D Protein Modeling and Protein-protein Interaction Prediction

Stella Chan Li Li (25158)

This progress report is submitted in partial fulfillment of the requirements for Final Year Project (STF3012)

Supervisor: Assoc. Prof. Dr. Edmund Sim

Programme of Resource Biotechnology

Department of Molecular Biology

Faculty of Resource Science and Technology

UNIVERSITI MALAYSIA SARAWAK

2012

Acknowledgement

This project would not have been possible without the support of many people. My heartfelt gratitude to my supervisor, Assoc. Prof. Dr. Edmund Sim who offered invaluable support and guidance. Special thanks also to Miss Ma Xiang Ru and friends from same final year project group for your care and attention. I could not thank you all enough.

DECLARATION

I hereby declare that no portion of the work referred in this project has been submitted in support of an application for another degree qualification of this or any other university or institution of higher learning.

(Stella Chan Li Li)
Resource Biotechnology
Department of Molecular Biology
Faculty of Resource Science and Technology
University Malaysia Sarawak

List of Abbreviations

3-D	3-Dimension
ARM	Armadillo repeat domain
DNA	Deoxyribonucleic acid
KOW	Kyprides, Ouzounis, Woese motif
MBT	Malignant Brain Tumor repeat
mRNA	Messenger Ribonucleic Acid
NPC	Nasopharyngeal Carcinoma
RMSD	Root mean square deviation
RNA	Ribonucleic Acid
RNP	Ribonucleic acid binding protein
RP	Ribosomal Protein
RPL	Ribosomal Protein Large subunit
rRNA	Ribosomal Ribonucleic Acid
SUMO	Small Ubiquitin-like Modifier
VAST	Vector Alignment Search Tool

List of Tables and Figures

Bioinformatics Analysis of the Ribosomal Proteins, RPL27, RPL37a and RPL41: 3-D Protein Modeling and Protein-protein Interaction Prediction

Stella Chan Li Li (25158)

Programme of Resource Biotechnology
Department of Molecular Biology
Faculty of Resource Science and Technology
UNIVERSITI MALAYSIA SARAWAK

ABSTRACT

Ribosomal proteins (RP) are constituents of ribosome that is important in protein biosynthesis and likely to have extraribosomal functions. Many RPs are related to various diseases and cancers. From previous studies, RPL27, RPL37a and RPL41 gene were reportedly downregulated in all cell lines derived from Nasopharyngeal Carcinoma (NPC) tissues compared to the normal counterpart. Thus, present study aimed to understand the three genes in protein level and its interaction with other proteins. The methods used RPL27, RPL37a and RPL41 3-D (3-Dimension) protein models to search for structural neighbor in predicting protein-protein interaction. Subsequently, the structural neighbors and their partners were docked to predict functions. As a result, RPL27 revealed interaction with SYNJ2 and UBC9 in dock complex. RPL27 was predicted to mediate RNA binding protein and deregulate sumoylation. Additionally, RPL37a interacted with CTNNB1, SCMH1 and ATBF1. RPL37a was predicted to deregulate Wnt degradation pathway and inhibit β-catenin migration. RPL37a might also regulate homeotic transcription. Further studies such as alanine scanning mutagenesis can provide deeper insight on protein recognition mechanism and identification of hot spots for protein kinetic studies.

Keywords: Ribosomal protein, RPL27, RPL37a, RPL41, comparative modeling, protein-protein interaction prediction.

ABSTRAK

Protein ribosom adalah sebahagian ribosom yang penting dalam biosintesis protein dan mungkin mempunyai fungsi-fungsi tambahan ribosom. Banyak protein ribosom dapat dikaitkan dengan pelbagai penyakit dan kanser. Kajian lepas melaporkan gen RPL27, RPL37a dan RPL41 bawah kawalan dalam sel daripada tisu-tisu karsinoma nasofarinks berbanding dengan yang biasa. Oleh itu, kajian ini bertujuan untuk memahami ketiga-tiga gen dalam peringkat protein dan interaksi dengan protein lain. Kaedah dalam kajian ini menggunakan RPL27, RPL37a dan RPL41 model protein 3-D untuk mencari jiran struktur dalam usaha meramalkan interaksi sesama protein. Seterusnya, jiran struktur dan calon rakan melalui docking untuk meramal fungsi protein. Hasilnya, RPL27 menunjukkan interaksi dengan SYNJ2 dan UBC9 dalam kompleks dock. RPL27 diramalkan mengawal protein pengikat RNA dan menyahkawal proses sumo. Selain itu, RPL37a berinteraksi dengan CTNNB1, SCMH1 dan ATBF1. RPL37a juga diramalkan menyahkawal laluan degradasi Wnt dan menghalang migrasi β-catenin. RPL37a juga kemungkinan mengawal transkripsi homeotik. Kajian lanjutan seperti mutagenesis pengimbasan alanine boleh memberikan penjelasan yang lebih mendalam dalam mekanisma pengiktirafan protein dan pengenalpastian 'hot spots' bagi kajian kinetik protein.

Kata kunci: *Protein ribosom, RPL27, RPL37a, RPL41, permodelan bandingan, ramalan interaksi sesama protein.*

1.0 Introduction

Ribosomal proteins (RP) are constituents of ribosome involved in protein biosynthesis. However, most RPs are structurally distinct, tissue specific and differ in regulation. Besides ribosome biosynthesis, several RPs possesses distinct extraribosomal functions and modulate function of important regulatory protein (Lindström, 2009). Meanwhile, disruptions in ribosome biogenesis are generally associated with diseases and cancer. From previous studies, RP genes showed association with colorectal carcinoma (Sim, Bong, Balraj, Tan, Jamal et al., 2006). A more recent study showed RPL27, RPL37a and RPL41 were significantly downregulated in all cell lines derived from Nasopharyngeal Carcinoma (NPC) compared to the normal counterpart (Sim, Ang, Ng, Lee & Narayanan, 2010). Therefore, present study aimed to understand the three genes in their protein level and its interaction with other proteins. Herein, 3-D protein models of these three proteins were constructed. This was in conjunction with experimental structures that are still unavailable and difficulties in obtaining ribosome components structures. The 3-D protein models served to search for structural neighbors in predicting protein-protein interactions. This approach was based on the assumption that structural neighbor and their partners interact similarly to RPL (Ribosomal Protein Large subunit) proteins. Subsequently, RPL models and their candidate partners were docked to predict functions from their putative interaction site. Docking was an amenable method due to difficulties in crystalizing transient complexes and modeling complexes. As a result, RPL27 and RPL37a were found interacting with a few proteins involved in important pathways. Nonetheless, further studies such as alanine scanning mutagenesis can be conducted on RPL dock complexes to identify individual amino acids responsible for protein recognition and identification of hot spots for protein kinetic studies.

The objectives of this study were:

1. To construct 3-D protein models of RPL27, RPL37a and RPL41

2. To predict protein interaction scenarios between RPL27, RPL37a, RPL41 and other proteins via structural neighbor approach and docking simulation.

3. To predict function of RPL27, RPL37a and RPL41 from dock complex.

2.0 Literature Review

2.1 Ribosomal proteins and extraribosomal functions

Ribosome consists of four ribosomal RNA (rRNA) species and 79 ribosomal proteins (RPs) in mammals (Wool, 1979). Ribosome evolution had RPs diverged into super families of proteins with individual characteristics. RPs are independent polypeptides and are structurally distinct. Their genes on the other hand showed tissue-specificity and was controlled by different regulators (Ishii, Washio, Uechi, Yoshihama, Kenmochi et al., 2006). Although varying in gene size and number of exons, human RP genes showed highly conserved sequences when compared to other organisms namely *Drosophila melagonaster, Caenorhabditis elegans* and *Saccharomyces cerevisae.*

The knowledge about number of RP genes sharing typical or unique features are still incomplete (Ishii et al., 2006). Nevertheless, RPs were found important in assembly of ribosomal subunits in protein biosynthesis but dispensable for function after ribonucleoprotein complex is fully assembled (Nierhaus, 1991). Some RPs were involved in other cellular processes besides protein biosynthesis. Several RPs possessed distinct extraribosomal functions in apoptosis, DNA repair and transcription. Meanwhile, there are increasing number of RPs shown to modulate trans-activation function of important regulatory protein such as NF-kB, p53, c-Myc and nuclear receptors (Lindström, 2009).

2.2 Ribosome biogenesis and Tumorigenesis

Ribosome biogenesis has been extensively studied in bacteria, yeast and flies, but less is known about ribosome biogenesis and RPs biology in mammals (Lindstrom, 2009). In general, ribosome biogenesis is the production and assembly of its rRNA molecules and RPs (Naora & Naora, 1999) in a series of concerted reactions (Warner & Nierras, 1998). Indeed, ribosome is a highly regulated complex involving regulatory transcription factor, rRNAs and ribosome associated proteins (PARs) (Lai & Xu, 2007).

Any increase or ceasing in ribosome biogenesis is usually associated to ribosomapathies (McGowan, Li, Park, Beaudry, Tabor et al., 2008). In many cases, it is related to cancers. Examples of RPs involved in human diseases are Diamond-Blackfan anaemia syndrome, Turner syndrome and Noonan syndrome (Liu, Li, Yu, Fu & Li, 2007).

RPs role in tumorigenesis was perturbated when ribosome biogenesis was disturbed or through extraribosomal functions. They can be linked to cell-cycle arrest or apoptosis. The disturbance in translation control was likely caused by deregulation of master regulators such as MYC and PTEN. As a result, components of translations are overexpressed. In another case, some RPs have extraribosomal function that regulates p53 and retinoblastoma (RB) to repress transcription process in tumorigenesis (Lai & Xu, 2007).

2.3 Ribosomal Protein and Cancer

Ribosomal proteins in cancer are usually seen in altered or unbalanced expression of RPs. In some cases, it has been used to distinguish normal cells from tumor cells. However so, it is not clear that RP expression is a by-product of increased cancer cell growth or a leading force in tumorigenesis (Lindström, 2009).

From previous studies, 33 RP genes showed consistent differential expression on Malaysian colorectal carcinoma in comparison with its normal tissues (Sim et al., 2006). The report postulated RPs association with cancer. In another relevant study, two out of ten genes (RPL26 and RPL27) (Sim, Toh & Tiong, 2008) followed by RPL37a and RPL41 (Sim et al., 2010) were identified downregulated in differential expression on Nasopharyngeal Carcinoma (NPC) compared to its normal counterpart. These findings suggested that RPs might play an important role in regulating NPC.

2.4 RPL27

RPL27 gene is located in Chr17q21.1-1.21.2. The 135 amino acids protein is a component of 60S subunit and belongs to L27E family. Its messenger RNA (mRNA) was found highly expressed in foetal development of kidney and several other foetal tissues (Gallagher, McClean & Malik, 1994).

RPL27 was postulated as candidate housekeeping gene for its stability in few expression studies (Jonge, Fehrmann, Bont, Hofstra, Gerbens et al., 2007; Lallemant, Evrard, Combescure, Chapuis, Chambon et al., 2009). In contrast, RPL27 also showed to have increased copy numbers and expression levels in microarray analysis of gastric carcinoma (Varis, Wolf, Monni, Vakkari, Kokkola et al., 2002). Moreover, RPL27 was suggested to have anti-cancer drug resistant phenotype when reported to highly inhibit in NPC-derived cell line (CNE2/DDP) (Jiang , Zhang, Yue, Zhu, Lu et al., 2003).

2.5 RPL37a

Full length human RPL37a clone was obtained from reverse transcription of partial clone homologous to rat RPL37a from human neuroblastoma cell line cDNA library (Saha, Tirumalai, Scala & Howells, 1993). Until recently, RPL37a gene was mapped to chromosome 2q35 based on an alignment of RPL37a sequence with genomic sequence (Hartz per. com. in Generation and initial analysis of more than 15,000 full- length human and mouse cDNA sequences, 2010). RPL37a protein belongs to L37AE family. The 92 amino acids protein is highly basic and contains putative zinc finger domain that may be involved in rRNA interactions (Saha et al., 1993).

Alike RPL27 for its stability, RPL37a was a reference gene for normalization of meningiomas in gene expression studies (Pfizer, Tatabiga & Rozer, 2011) and was associated to glioblastoma (Serao, Delfino, Southey, Beever & Rodriguwz-Zas, 2011).

2.6 RPL41

The coding sequence size of RPL41 is shortest among other RPs. RPL41 is only 25 amino acids long but highly specific (Ishii et al., 2006). RPL41 was highly expressed in skeletal muscle (Bortoluzzi, d'Alessi, Romualdi & Danieli, 2001).

RPL41 is important in mitosis as RPL41-depleted cells had premature splitting of centrosome while gene knockout showed abnormal spindles, frequent failure of cytokinesis and formation of polynuclear cells which all can be related to malignant transformation (Wang, Huang, He, Wang et al., 2010). In addition, RPL41 was also detected in 59% of tumor cell lines by fluorescence *in situ* hybridization (FISH) analysis and downregulation in 75% of primary breast cancers by reverse transcriptase polymerase chain reaction (Wang et al., 2010).

Furthermore, RPL41 protein was found to interact with β-subunit of protein kinase CKII and stimulated phosphorylation of DNA topoisomerase-α (Lee, Hwang, Chen, Lin, Huang et al., 2007). RPL41 was also a candidate gene with SFR4 in normalization of hepatitis C virus (Waxman & Wumbach, 2007). Besides that, RPL41 induced rapid degradation of ATF4, a transcription factor critical for tumor cell survival in stress (Wang, Xu, Zhang, He, Ya et al., 2011).

2.7 Protein Modeling

Recently, RP studies in correlation with disease and cancers are directed to translational level (Ruggero & Pandolfi, 2003). However, the three proteins in this study still have no experimental structures. Due to the inherently dynamic behavior of ribosome, many transient structures are difficult to trap and crystallize (Schmeing & Ramakrishnan, 2009). Besides that, many of the RP structures cannot be characterized from the subunit structures when isolated. The globular domains with long extensions and tails could hardly have a stable structure in isolation (Klein, Moore & Steitz, 2004).

Experimental methods for structure determination such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) are time-consuming and limited in application thus unable to keep pace with flood of newly characterized gene products (Al-lazkani, Jung, Xiang & Honig, 2001). However, *in silico* protein modeling is an alternative to predict structure.

In the present study, comparative modeling of the three RPL proteins was conducted. This approach can predict 3-D structure with accuracy comparable to that of low resolution experimental structures (Sanchez & Sali, 1997).

2.8 Protein-protein interaction prediction

Studies on protein-protein interaction are increasing due to the importance in understanding their holistic role within the cell. Indeed, proteins do not function in isolation but with other macromolecules. Protein interaction that are disrupted and resulted in change of stability or inhibition in binding to other molecules is the causes of many diseases. Also, proteins tend to interact with other proteins involved in the same disease (Kann, 2007).

Interaction studies with protein of unknown function as target, several clusters can be extracted from neighboring proteins based on their structural similarity (Monji, Koizumi, Ozaki & Ohkawa, 2011). However, experimental determination of protein-protein interaction in complexes is expensive and time-consuming as well as problematic for transient complexes, while the prediction of complexes by comparative modeling is suitable in relatively few cases.

Therefore, an alternative is protein-protein docking. Docking fit two or more known structures or reliable 3-D structural models via their interacting surfaces. It is an attractive alternative especially when proteins in complexes are still under-represented in the structural databases (Ezkurdia, Bartoli, Fariselli, Casadio et al., 2009).

2.9 Bioinformatic Analysis

2.9.1 Clustal X

The Clustal series of programs are widely used for multiple alignments of both nucleic acids and protein sequences. The programs are well known for accuracy of results, robustness, portability and user-friendliness. Clustal X is the newer windows interface for Clustal W. The default options of Clustal W and Clustal X 2.0 are the same as Clustal W 1.83. They give the same alignment results. The integrated system has versatile sequence coloring scheme, pull-down menus, ability to cut-and-paste sequences, selection of sequence subset to realign and insert back to original alignment. Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted (Thompson, Gibson, Plewniak, Jeanmougin et al., 1997). The latest ClustalX 2.0 has been completely rewritten in C++ language for proper porting of program to latest version of computer operating systems (Larkin, Blackshields, Brown, Chenna et al., 2007). Clustal X is a standalone program.

2.9.2 SWISS-MODEL

SWISS-MODEL workspace is a web-based integrated service for automated comparative modeling of 3-D protein structures (Schwede, Koop, Guex & Peitsch, 2003). It is developed by BIOZENTRUM laboratory and Swiss Institute of Bioinformatics (SIB).

Currently, SWISS-MODEL is the most widely used modeling facility. It can be used at all stages of comparative modeling including identification of structural template(s), alignment of target sequences and template structure(s), model building and model quality evaluation. Moreover, these steps can be repeated to achieve a satisfactory modeling result. The workspace extended the web-based system by assisting manual user intervention in model building and evaluation (Arnold, Bordoli, Kopp & Schwede, 2006). The SWISS-MODEL workspace can be accessed freely at http://swissmodel.expasy.org/workspace/

2.9.3 VAST

Vector Alignment Search Tool (VAST) is a computer algorithm to identify protein 3-D structure by similarity and distant evolutionary relationships. This tool can be used with any protein molecule of interest already in Protein Data Bank (PDB) or input protein molecule in PDB format. The search is against Molecular Modeling Database (MMDB) which are experimentally resolved structures derived from PDB. VAST neighbors include only those neighbors classified as belonging to the same homologous superfamily by SCOP or 12% sequence identity in VAST structure alignment (Madej, Panchenko, Chen & Bryant, 2007). Using structural similarity as query and database proteins as standard of truth, position-specific score matrices (PSSMs) recognition sensitivity depends on primarily on the diversity of sequences included in the alignment, with an optimum around 30-50% average pairwise identity (Panchenko & Madej, 2004). VAST can be accessed from www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html.

2.9.4 IntAct

IntAct is molecular interaction database found by IMEx Consortium, a collaboration of interaction databases. It is an open source and its data are curated from literature or from direct data depositions. As at September 2011, IntAct database contains approximately 275,000 curated binary interactions evidences from over 5000 publications. The information within the database primarily consisted of protein-protein interaction data. The improved database has enhanced search process and new data formats (Kerrien, Aranda, Breuza, Bridge et al., 2011). The database can be reached at http://www/ebi.ac.uk/intact.

2.9.5 ClusPro

ClusPro represents the first fully automated, web-based program for computational docking of protein structures. Two protein structures can be uploaded in PDB format through ClusPro web interface or simply PDB codes where ClusPro download from PDB servers. ClusPro algorithms are such as DOT or ZDOCK which perform rigid body docking based on fast Fourier transform (FFT) correlation techniques. Filtering via empirical free energy functions is then applied to set of structures to select those with good electrostatic and desolvation free energies for further pairewise root mean square deviation (RMSD) clustering (Comeau, Gatchell, Vajda & Camacho, 2004). ClusPro 2.0 was found to have the best performance in latest round of Critical Assessment of Predicted Interactions (CAPRI). The ClusPro 2.0 server is available at http://cluspro.bu.edu.

3.0 Materials and Methods

A flow chart of methods for constructing 3-D protein model and predict protein-protein interaction as well as functions of RPL27, RPL37a and RPL41 were shown in figure 1. The methods involved target sequence from Ribosomal Protein Gene (RPG) database (Nakao, Yoshihama & Kenmochi, 2004) and template identification using PSI-BLAST (Altschul, Madden, Schäffer, Zhang & Zhang, 1997), followed by target-template alignments using ClustalX. Then, method continued with 3-D protein modeling using SWISS-MODEL and model evaluation using QMEAN6 scoring (Benkert, Kunzli & Schwede, 2009) and PROCHECK (Laskowski, MacArthur, Moss &Thornton, 1993) analysis. Later, structural neighbors of the model were identified using VAST while their respective partners from IntAct database. This was followed by docking of the model with candidate partners using ClusPro. The dock complex was used to predict protein-protein interaction scenarios and RPL protein functions. The detailed descriptions of the methods were as followed.

3.1 Target sequences, template identification and selection

Gene and protein sequences of human RPL27, RPL37a and RPL41 were retrieved from RPG database (Table 1). A PSI-BLAST search for each query RPL sequences against PDB structures was performed. The search excluded model and metagenomic sequences. Due to short sequence of RPL41 (25 amino acids), parameter was altered to expected threshold of 20000, word size reduced to two, no compositional adjustment and PAM30 protein matrix.

Template structure with E-value above threshold with reasonable percentage of sequence identity and resolution were selected for multiple sequence alignment.

Table 1: Gene and protein ID of RPL27, RPL37a and RPL41.

Gene	Gene ID (Accession no.)	Protein ID (Accession no.)
RPL27	AC055866	NP_000979
RPL37a	AC073321	NP_000989
RPL41	AB062066 D28462	NP_001030344.1 NP_066927.1

3.2 Target –template alignments

The selected template sequences were aligned to target sequence via ClustalX. Protein weight matrix used was of BLOSUM series while clustering algorithm used was UPGMA algorithm. Each alignment step was iterated. Every iteration effectively refined alignment by removing and realigning each sequence.