DoChaP: The Domain Change Presenter

Shani T. Gal-Oz¹, Nimrod Haiat², Dana Eliyahu², Guy Shani², Tal Shay^{1,2*}

¹ Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva,
Israel

² Department of Software and Information Systems Engineering, Ben-Gurion
University of the Negev, Beer-Sheva, Israel

*corresponding author: Tal Shay, PhD. Email: talshay@bgu.ac.il

Abstract

Alternative RNA splicing results in multiple transcripts of the same gene, possibly encoding for different protein isoforms with different protein domains and functionalities. Whereas it is possible to manually determine the effect of a specific alternative splicing event on the domain composition of a particular encoded protein, the process requires the tedious integration of several data sources; it is therefore error prone and its implementation is not feasible for genome-wide characterization of domains affected by differential splicing. To fulfill the need for an automated solution, we developed the Domain Change Presenter (DoChaP), a web server for the visualization of the exon–domain association. DoChaP visualizes all transcripts of a given gene, the domains of the proteins that they encode, and the exons encoding each domain. The visualization enables a comparison between the transcripts and between the protein isoforms they encode for. The organization and visual presentation of the information makes the structural effect of each alternative splicing event on the protein structure easily identified. To enable a study of the conservation of the exon structure, alternative splicing, and the effect of alternative splicing on protein domains, DoChaP also facilitates an inter-species comparison of domain–exon associations. DoChaP thus provides a unique and easy-to-use visualization of the exon–domain association and its conservation between transcripts and orthologous genes and will facilitate the study of the functional effects of alternative splicing in health and disease.

Introduction

Alternative splicing creates multiple transcripts from a single gene, possibly encoding for different protein isoforms, and therefore increases the functional complexity of the eukaryotic proteome¹⁻⁴. Proteins are composed of structural domains, each of which may have an independent function⁵. Protein domains can be encoded entirely from one exon or from several consecutive exons. In both cases, alternative splicing can cause either the loss of the domain or a change in the length and composition of the RNA sequence that codes for the domain and hence in the functionality of the domain or even of the entire protein. Indeed, it has previously been reported that alternative splicing changes protein functionality in tissue-specific splicing⁶, protein interactions⁷, cancer^{8,9}, resistance to biological therapeutics¹⁰, and the immune response to viral infections¹¹.

Box 1. Manual determination of the effect of alternative splicing on protein domains in a single gene

1. Identify the transcripts of the gene (from RefSeq and/or Ensembl)
2. Find the protein isoforms that are encoded by the identified transcripts (from RefSeq and/or Ensembl)
3. Predict the domains in each isoform separately (e.g., by Pfam or SMART) or look for the domains of the isoform in databases (e.g., NCBI's CDD)
4. Manually compare the type, position, and length of the protein domains found between each pair of isoforms.
5. For each domain whose properties have changed, translate the positions of the domain in the two protein isoforms into positions in the transcripts, translate positions in the transcripts into genomic positions, and examine which exon(s) of the transcripts that encode the pair of isoforms are in this position and how this(these) exon(s) has(have) changed between the transcripts.
6. For each domain that is missing in one of the isoforms, translate the position in the protein that carries the domain into the position in the transcript, translate the position in the transcript into the genomic position, and examine which exon is in this position and how this exon has changed between the transcripts.

The use of high throughput RNA sequencing (RNA-seq) data to study genome wide alternative and differential splicing is becoming more and more common¹². However, determining the functional implications of alternative splicing at the protein level involves a tedious manual process (Box 1) for each case of alternative splicing. Furthermore, experimental biologists who are studying specific cases of different functionalities of alternative protein isoforms currently do not have a tool for exploring the different isoforms of the protein alongside their domains.

Currently available domain prediction tools (e.g., SMART¹³, Pfam¹⁴, NCBI's CDD¹⁵, and TIGRFAMs¹⁶) predict protein domains either from the amino acid sequence of known proteins or from user supplied query sequences. Such tools provide the domains' annotation and visualization for the protein of interest, and some even specify the location of the introns¹³, but they can present only a single protein isoform (encoded by a single transcript) at a time, regardless of the other known isoforms and their domain compositions. Several database tools have thus attempted to address the problem of studying the effect of alternative splicing on protein domains, e.g., ExDom¹⁷, ProSAS¹⁸, and ASPicDB¹⁹. However, ExDom¹⁷ and ProSaS¹⁸ were built around a decade ago on much smaller source databases, and at present they are down and no longer maintained. ASPicDB¹⁹ is limited to human genome and does not show the transcript to isoform association or the exon to domain association. Therefore, the gap between the amount of existing information about the effect of alternative splicing on protein domains and the ability of researchers to visualize this effect remains to be breached.

To close this gap and thereby to provide researchers with an intuitive visualization of alternative splicing and information on protein domains, we built the Domain Change Presenter (DoChaP) web server. DoChaP provides a user-friendly, simple and intuitive gene-centric visualization of all the alternative splicing isoforms of a gene and the

domains that they encode (Figure 1). DoChaP covers five species and presents the genes in their genomic context, alongside their transcripts and protein domains, thereby highlighting the potential connections between exons and the protein domains that they encode.

Methods

Data Sources

DoChaP integrates information for transcripts and protein domains from several sources (Table 1) into a single SQLite database (Figure 1A). RefSeq²⁰ and Ensembl's²¹ transcripts, gene CDSs and exon annotations were extracted from GFF (general feature format) files, downloaded directly from the ftp sites of NCBI's Genome (<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>) and Ensembl (ftp://ftp.ensembl.org/pub/current_gff3/). Protein information, connections between transcripts and protein isoforms, and domain annotations, descriptions and external identifiers were taken from: (1) RefSeq: GenPept flat files (.gpff), downloaded from NCBI's RefSeq ftp site (ftp://ftp.ncbi.nlm.nih.gov/refseq/{species-specific}/mRNA_Prot/); (2) Ensembl: BioMart data mining tool²² (<https://m.ensembl.org/biomart/martview/>), using xml queries (templates can be found in the GitHub repository; Figure 1B).

RefSeq and Ensembl use domain predictions from multiple sources, of which DoChaP presents only NCBI's Conserved Domains Database (CDD)¹⁵, Pfam¹⁴, SMART¹³, TIGRFAMs¹⁶ and InterPro²³. For DoChaP, the connection between domain accessions from different sources was inferred from InterPro entries that were downloaded directly from InterPro²³ (<https://www.ebi.ac.uk/interpro/entry/InterPro/#table>). The association between RefSeq and Ensembl accession numbers was taken from the Gene2ensembl

table downloaded from NCBI's Gene ftp site (<ftp://ftp.ncbi.nih.gov/gene/DATA/> ; Figure 1B).

Orthology data for each pair of species in the DoChaP database was downloaded from Ensembl Compara via the BioMart data mining tool, using xml queries (xml templates can be found in the GitHub repository; Figure 1B).

Table 1. DoChaP sources

Source	Download path	File/table names	Obtained information
NCBI's Genomes	https://ftp.ncbi.nlm.nih.gov/genomes/refseq/	Gff file	RefSeq's transcripts, genes CDS and exons annotations
NCBI's RefSeq	ftp://ftp.ncbi.nih.gov/refseq/ <species> /mRNA_Prot/	All GPF files per species	RefSeq's proteins and protein domains data
Ensembl	ftp://ftp.ensembl.org/pub/current_gff3/	Gff3 file	Ensembl's transcripts, genes CDS and exons annotations
Ensembl	https://m.ensembl.org/biomart/martview/	Xml query templates are to be found in the GitHub	Ensembl's protein domains data Orthology information
InterPro	https://www.ebi.ac.uk/interpro/entry/InterPro/#table	Entries_table	Connection between domains accession from different sources
NCBI's Gene	ftp://ftp.ncbi.nih.gov/gene/DATA/	Gene2ensembl	Transition from RefSeq to Ensembl identifiers

Figure 1. Overview. (A) The DoChaP-database (DoChaP-db) integrates information on transcripts and protein domains from several sources into a single SQLite database. (B) Genes, transcripts, exons, proteins and domain events data was taken from NCBI and Ensembl. Orthology information was taken from Ensembl Compara. Domain types and the mapping of the domain accessions from different sources (Pfam, SMART, CDD, and TIGRFAMs) was inferred from InterPro. (C) A simple visualization of the information collected into the DoChaP-db is presented in the DoChaP webserver for a user selected gene. The visualization presents the genomic context of the transcripts (Gene) and the exon structure of the coding region of the mature mRNA of all the transcripts (T1, T2) and their encoded protein domain composition (P1, P2). Exons and their encoded protein domains are represented in different colors. (D) The 'species comparison' feature allows the user to select a gene and a species and visualize the orthologous genes so as to enable a comparison of transcripts and their protein domain organization.

Database content

The DoChaP database currently includes information for five species: human (*Homo sapiens*, hg38), mouse (*Mus musculus*, mm10), rat (*Rattus norvegicus*, rn6), zebrafish (*Danio rerio*, danRer11) and frog (*Xenopus tropicalis*, xenTro9). The database content is detailed in Table 2 (relevant to October 2020), and the database schema is detailed in Supplementary Figure 1.

Implementation

The DoChaP database was built using Python 3. The Ftplib package (<https://docs.python.org/3/library/ftplib.html>) was used for downloading from the ftp sites. The Bio.SeqIO module of the Biopython²⁴ package was used for parsing the GFF and GPFF; the pandas²⁵ package was used for data parsing; and the SQLite3 package (<https://docs.python.org/3/library/sqlite3.html>) was used for building the database.

The DoChaP web server is implemented with an Express.js on Node.js environment and is distributed via XAMPP. The website is controlled by AngularJS and uses HTML5 canvas for visualization.

Table 2. DoChaP database content for five species

Species	Gene IDs	Transcript - isoform pairs	Unique exons	Protein domains	Domain occurrences	In-domain junction
<i>Homo sapiens</i>	20,388	173,783	420,565	10,855	366,414	819,397
<i>Mus musculus</i>	22,971	118,560	351,417	10,352	267,835	581,283
<i>Rattus norvegicus</i>	25,377	63,812	287,556	10,507	151,469	352,689
<i>Danio rerio</i>	29,538	70,105	352,955	10,165	264,619	547,192
<i>Xenopus tropicalis</i>	31,912	88,166	473,954	9,277	331,223	938,047
Total	130,186	514,426	1,886,447	51,156	1,381,560	3,238,608

Results

Features and input

The DoChaP web server currently includes two visualization options for the exon–domain relationship: (1) The 'single species transcript comparison' provides genomic, transcript and domain composition visualization of all the known RefSeq and Ensembl transcripts of a gene of interest in a specified species (Figure 1C). (2) The 'species

comparison' provides the same information for a gene of interest in a selected species and its orthologous gene in one of the other available species (Figure 1D).

The search term for DoChaP can be in any of the following formats: gene, transcript or protein identifiers (gene symbol; NCBI's Gene ID or RefSeq transcript or protein ID; Ensembl gene, transcript or protein ID). In the single species transcript comparison, the species of interest should be specified. In the species comparison feature, the user should select a gene and species of interest, DoChaP will then search for all the orthologs of the gene in the other species and allow the user to select which species and orthologs to present. Orthologous genes are taken from Ensembl Compara. As some genes with identical gene symbols from different species are not defined as orthologs in Ensembl Compara (e.g., *A2ml1* in human and mouse), comparison of such genes is also enabled in DoChaP.

Graphical Visualization

DoChaP provides an intuitive, informative and straight-forward way of exploring the annotations of exons and domains. In the results page, all the transcripts of the gene of interest are shown with the following elements:

- a. Genomic visualization: The genomic visualization provides the genomic context of the shown transcript, including the loci, strand direction, and introns, exons and coding region. Exons are represented in different colors that are consistent across all the transcripts of the gene and through all the visualizations. Information on the start and end positions of the exon and its ordinal number in the transcript is available by mouse hovering on each exon. Each transcript can be manually hidden from the view by clicking the "hide transcript" button (another click will show it). The user can adjust the genomic range shown so as to focus on a specific set of exons. For genes that are coded on the reverse strand, the genomic visualization will

be reversed (horizontally flipped) so that, regardless of the coding strand, the location of first exon is the left-most and that of the last exon is the right-most.

- b. Transcript visualization: For each transcript in the genomic visualization, the exon structure of the coding region of the mature mRNA (without the untranslated regions) is provided. Exons are represented in different colors, corresponding to the relevant colors in the genomic visualization and being consistent across transcripts. The length and ordinal number of the exon in the transcript are available by mouse hovering over the exon. Transcript RefSeq and/or Ensembl IDs are shown above each transcript. The nucleotide and amino acid length scales are shown at the top. The user can adjust the genomic range displayed so as to focus on a specific set of exons and domains.
- c. Protein composition visualization: Protein domains are shown on a black line below the mRNA exon structure, and are represented with ellipses. The colors and positions of the domains are consistent with the encoding exon or exons. Different colors within the same domain indicate that more than one exon codes for the domain (in-domain junction). The domain name is shown below the domain rounded shape. By mouse hovering over the domain, additional details such as the full name, position on the protein, length in amino acids and link to the external domain source are displayed. For simplicity, overlapping domains are grouped and represented as a single ellipse with a double line. Mouse hovering will show how many domains are represented within this shape and a click on it will expand the visualization to show all the grouped overlapping domains. Protein RefSeq and/or Ensembl IDs are shown above each protein isoform visualization.
- d. Color guide exons table: This table is shown on the bottom of the results page. Each color assigned to an exon in that page is shown, and the start and end positions as

well as the IDs of all the transcripts in which the exon is included are detailed in the table.

The 'change display option' allows the user to choose which types of visualization are shown: genomic only, transcript only, protein composition only or transcript and protein composition together. The default view displays all views. The 'hide predicted records' option enables removal of all RefSeq's predicted transcripts and proteins (prefix XM_ and XP_), and only curated RefSeq (prefix NM_ and NP) and Ensembl (prefix ENS) records will be displayed in the results. In the species comparison tab, all transcripts of the relevant gene from both species are shown in similar format, side by side. Each result page can be exported to a PDF by using the 'Save as PDF' button.

Example and Biological impact

To demonstrate how the visualization of DoChaP can promote the understanding of the functional effect of alternative splicing, we use the tumor suppressor breast cancer gene 1 (*BRCA1*) as a model gene. *BRCA1* is one of the most commonly mutated genes in breast and ovarian cancer²⁶. In Figure 2, we present the DoChaP view of human *BRCA1* (4 representative transcripts out of 32). *BRCA1* is encoded on the reverse DNA strand and consists of 24 exons (Figure 2A). Alternative splice variants of *BRCA1*, mainly focused in exons 1, 3 and 10 (E1a/b, E3, E10, respectively; Figure 2B and C), have been previously reported and widely studied in different contexts²⁷⁻²⁹. The frequency of the alternative transcripts of *BRCA1* changes between tissues²⁹, and also might play a role in the tumorigenesis of breast cancer³⁰. For example, exon number 3, is skipped in *BRCA1* transcript NM_007297 (E3, Figure 2A left and zoom in Figure 2B). Since this exon encodes part of the RING finger motif (Figure 2A, right panel), the resulting protein product lacks the RING finger domain ("Znf RING"), and this, in turn, might affect the cell cycle regulatory function of this protein isoform²⁹. The largest exon of

the *BRCA1* gene is exon E10 (3,426 bp; Figure 2A left and zoom in Figure 2C, light blue). Strangely, this exon is sometimes referred to as exon 11, because of an upstream Alu element that was inserted into BRCA1 clone (historically referred to mistakenly as exon 4)²⁶. Two of the presented transcripts (NM_007297 and NM_007300) include the full-length E10 (3,426 bp, Figure 2c, light blue). However, in transcripts NM_007298 and NM_007299, an alternative 5' splice site is used and only the first 117 bp of this exon (E10₁₁₇) remain in the mature mRNA (Figure 2c, purple). The alternative 5' splice site of exon 10 causes the loss of specific protein binding sites [such as RB, p53, MYC, RAD50, TUBG (γ -tubulin) and angiopoietin-1] and the loss of a nuclear localization signal, and therefore affects the functionality of the translated protein²⁷. The “BRCA1 serine dom” domain, which is a serine-rich domain associated with the BRCA1 C-terminus, is encoded from the full version of exon 10 (Figure 2a, right panel, light blue) and is seen only in the protein products of the transcripts that include the full-length exon 10 (Figure 2a, right panel). In the case of *BRCA1*, DoChAP provides a clear and intuitive visualization of the complex pieces of information that have been collected over many years in several different studies and of all the known domains that exist (or not) according to the annotation of all transcripts.

Limitations

The public databases used as sources for DoChaP include redundancies and inconsistencies. To overcome this problem, we currently use the RefSeq – Ensembl conversion table `gene2ensembl` (from NCBI's ftp, see methods). In cases where matched RefSeq and Ensembl protein records have different lengths (of more than one amino acid; for example, the protein isoform NP_001269863.2 and ENSP00000408592.1 of the gene *Prdm15* have different lengths of 1,161 and 1,198 amino acids, respectively), the records were separated into two transcripts in the visualization. In all other cases, matched records were merged into one transcript to avoid redundancies.

Different data sources may indicate slightly different start and end positions of the domains in the protein, probably due to differences in domain prediction methods [for example, in the protein isoform NP_000236.2 of the human gene *Met*, pfam has located the Sema domain between amino acids 57-490 (pfam01403) and SMART has located the same domain in 52-496 (smart00630)]. In such cases, we present the coordinates of all the overlapping domains and show the source from which the coordinates of each domain occurrence were taken. Overlapping domains are shown as one ellipse with inner circle line that can be expanded upon click.

Some very long transcripts and transcripts with many exons make the genomic visualization of an entire gene less efficient in terms of colors and relative exon sizes. In such cases, the user is advised to zoom in and explore the gene in several consecutive windows.

Conclusions and Future Extensions

Improved sequencing technologies are leading to increased interest in the connection between exons and domains and the evolution of the regulatory mechanisms of alternative and differential splicing. Previously, to explore exon–domain associations, researchers had to manually search, integrate and interpret textual data from several sources in a tedious and error-prone process. With the development of DoChaP (made possible by the increasing quantities of freely available data in the public domain), the research community has access to a fast, intuitive and easy-to-use visualization tool for exploring the exon–domain relationship.

DoChaP is an ongoing project, and we plan to add new features and visualization capabilities and increase the database in updates. Specifically, we intend to add additional species and other types of functional features of proteins, such as signal peptides. Moreover, we will integrate a method for comparison of transcripts that will allow DoChaP to order the transcripts based on their genomic similarity and domain organization and to present similarity scores for transcripts of the same gene and between orthologs.

Given the increased interest in the effect of alternative splicing on functionality in health and disease, we believe that DoChaP will allow researchers from the biological, medical and computational fields to interpret the structural effect of their transcriptional profiling changes toward new functional findings.

Data Availability

The DoChaP web server is freely available at <https://DoChaP.bgu.ac.il>. The latest DoChaP database may be downloaded under “Downloads” tab in the DoChaP website. The source code is available on GitHub (<https://github.com/Tal-Shay-Group/DoChaP>).

Acknowledgements

The authors would like to thank Yehuda Barouch for his continuous valuable support during DoChaP server development, and Eran Lachs and Yossi Gross from the Technologies, Innovation & Digital Division at Ben-Gurion University for the technical support during server setup.

Funding

Research reported in this publication was supported by Israel Science Foundation Grant 500/15 and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R24AI072073. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. S.T.G.-O. was supported by Hi-Tech, Bio-Tech, and Chemo-tech fellowship and the Negev fellowship of Ben-Gurion University of the Negev.

References

1. Irimia, M., Rukov, J. L., Penny, D. & Roy, S. W. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC evolutionary biology* 7, 188 (2007).
2. Ast, G. How did alternative splicing evolve? *Nature reviews. Genetics* 5, 773-782 (2004).
3. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001).
4. Liu, Y. et al. Impact of Alternative Splicing on the Human Proteome. *Cell Reports* 20, 1229-1241 (2017).
5. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* 300, 1701-1703 (2003).
6. Buljan, M. et al. Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular cell* 46, 871-883 (2012).
7. Yang, X. et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* 164, 805-817 (2016).
8. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell reports* 20, 2215-2226 (2017).
9. Au, W. W. Y. & Henderson, B. R. The BRCA1 RING and BRCT Domains Cooperate in Targeting BRCA1 to Ionizing Radiation-induced Nuclear Foci. *J. Biol. Chem.* 280, 6993-7001 (2005).
10. Poulikakos, P. I. et al. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature* 480, 387-390 (2011).
11. Shemer-Avni, Y. et al. Expression of NKp46 Splice Variants in Nasal Lavage Following Respiratory Viral Infection: Domain 1-Negative Isoforms Predominate and Manifest Higher Activity. *Front. Immunol.* 8 (2017).
12. Mehmood, A. et al. Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief Bioinform* 21, 2052-2065 (2020).
13. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493-D496 (2018).
14. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427-D432 (2019).
15. Lu, S. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265-D268 (2020).

16. Haft, D. H. et al. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* 41, 387 (2013).
17. Bhasi, A., Philip, P., Manikandan, V. & Senapathy, P. ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.* 37, 703 (2009).
18. Birzele, F. et al. ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.* 36, 63 (2008).
19. D'Antonio, M. et al. ASPicDB: a database web tool for alternative splicing analysis. *Methods Mol. Biol.* 1269, 365-378 (2015).
20. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, 733 (2016).
21. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* 48, D682-D688 (2020).
22. Kinsella, R. J. et al. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database* 2011, bar030 (2011).
23. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351-D360 (2019).
24. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422-1423 (2009).
25. McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61 (2010).
26. Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266, 66-71 (1994).
27. Li, D., Harlan-Williams, L. M., Kumaraswamy, E. & Jensen, R. A. BRCA1—No Matter How You Splice It. *Cancer Res.* 79, 2091-2098 (2019).
28. Xiping, Z., Qingshan, W., Shuai, Z., Hongjian, Y. & Xiaowen, D. A summary of relationships between alternative splicing and breast cancer. *Oncotarget* 8, 51986-51993 (2017).
29. Orban, T. I. & Olah, E. Emerging roles of BRCA1 alternative splicing. *Mol. Pathol.* 56, 191-197 (2003).
30. Lixia, M., Zhijian, C., Chao, S., Chaojiang, G. & Congyi, Z. Alternative splicing of breast cancer associated gene BRCA1 from breast cancer cell line. *J. Biochem. Mol. Biol.* 40, 15-21 (2007).