Effects of Precipitation Methods on the Properties of Protease Extracted from Starfruit (*Averrhoa carambola* L.) of Different Maturity Index

Normah Ismail and Ezzana Zuraini Zainuddin

Department of Food Technology, Faculty of Applied Sciences,
Universiti Teknologi MARA (UiTM),
40450 Shah Alam, Selangor Darul Ehsan
E-mail: norismel@salam.uitm.edu.my

ABSTRACT

Proteases were extracted from starfruit at maturity Index 2 (unripe, light green) and Index 7 (very ripe, orange) and partially purified using acetone and 40% ammonium sulfate precipitations. Higher yield and proteolytic activity were observed for proteases purified using acetone than 40% ammonium sulfate. As for maturity index, yield and protein concentration of proteases from Index 2 were higher than those from Index 7. SDS-PAGE result showed intense bands for acetone proteases while a distinct band at 50 kDa was observed in all the proteases. Enzyme activity decreased during the seven days storage at 4°C with minimum relative activity of 70% achieved for acetone proteases at day seven. This study suggested that acetone precipitation is more effective method for purifying starfruit protease based on the yield and proteolytic activity compared to using 40% ammonium sulphate precipitation. In order to obtain higher protein concentration and proteolytic activity, starfruit at the unripe stage, Index 2 is a better raw material than Index 7 to be used for protease production.

Keywords: *starfruit* (*Averrhoa carambola L.*)*, protease, purification, acetone, ammonium sulfate*

INTRODUCTION

Starfruit (*Averrhoa carambola L.*), categorized under Oxalidaceae family, is one of the widely grown tropical fruits in Malaysia especially in Selangor, Negeri Sembilan and Johor. The fruit which is sweet and slightly acidic, succulent and juicy with attractive flesh and distinctive flavor is usually eaten fresh, also served as fresh juices or used as flavor ingredients in juice blends [1].

Proteases are enzymes that breakdown protein. They are classified according to their sources (i.e., animal, plant, microbial), catalytic action (i.e., endopeptidase or exopeptidase) and nature of the catalytic site [2]. Some fruits have already been known to contain high amount of protease, for example in young fruit of papaya (*Carica papaya*); the protease is abundantly found in the latex in the form of papain, chymopapain and papaya peptidase A [3]. Protease is also found in fig (ficin) as well as fruit and stem of pineapple (bromelain) [4]. Proteases are routinely used in cheese making, baking and meat tenderization. Most plant proteases are active over a wide range of pH.

Protease has been purified by several methods including salt precipitation and chromatography [5], three-phase partitioning (TPP) [6], extraction by homogenizing in Tris-HCl buffer followed by purification in ammonium sulfate [7, 8] and extraction using phosphate buffer and subsequent purification with acetone precipitation [9, 10]. In this study, proteases from the unripe and ripe starfruit were extracted, purified using acetone and ammonium sulfate precipitation and the effects of the purification methods on the proteolytic activity, molecular weight distribution and storage stability of the extracted and purified proteases were determined.

MATERIALS AND METHOD

Plant Materials and Chemicals

Starfruit was purchased from Malaysian Agricultural Research and Development Institute (MARDI), Jelebu, Negeri Sembilan. Starfruits with maturity indices 2 (unripe, light green) and 7 (very ripe, orange) were used in this study. All chemicals and reagents used were of analytical grade.

Extraction of Protease from Starfruit

The fruit was cut and the seeds were removed before being ground in a juice extractor. The juice was then filtered through a layer of muslin cloth and stored at 4°C.

Purification of Protease

Two different purification methods were used to purify the crude extract comprising of acetone and ammonium sulfate precipitation. Acetone precipitation was performed according to the method of He *et al.* [9]. Cold acetone (-20^{o}C) was slowly added into the crude extract and the mixture was gently agitated to allow precipitation. This was followed by centrifugation using a centrifuge (Model 5420 Kubota, Japan) at 10,000 rpm for 15 min. The precipitate was then dissolved in phosphate buffer (50 mM), pH 7.2 and dialyzed. Ammonium sulfate precipitation was performed according to Wang *et al.* [7]. One hundred milliliter crude extract was mixed with ammonium sulfate to a concentration of 40% (w/v) followed by 4 hrs incubation at 4°C to precipitate the protease. This was followed by centrifugation at 10,000 RCF for 10 min using a centrifuge (Model 5420 Kubota, Japan). The precipitate was collected and then dissolved in 0.02 M Tris-HCI buffer, pH 7.5 and dialyzed at 4°C for 12 hrs.

Protein Content and Protein Concentration

Protein content in starfruit was determined using Kjeldahl method [11]. Protein concentration was determined using Bradford method [12].

Total Activity of Proteases

Proteolytic activity assay was performed according to the method of Kaneda & Uchikoba [13] with slight modification. Protease at 0.1 mL was added into 0.9 mL of 1% (w/v) casein dissolved in 0.2M sodium phosphate buffer solution at pH 7. The mixture was incubated at 38^{o} C for 20 min. 3 mL trichloroacetic acid (5% w/v) is added. After 30 min, the precipitate was removed by centrifugation at 10,000 RCF for 20 min using a centrifuge (Model 5420, Kubota, Japan). The absorbance of the supernatant was measured at 280 nm using UV-vis spectrophotometer.

Protease Specific Activity

Protease unit per ml divided by protein in mg/ml concentration indicated the specific protease activity [14] where:

$$\text{Specific activity (CDU/mg)} = \frac{\text{enzyme units (CDU)/ml}}{\text{protein in mg/ml}}$$

Molecular Weight Distribution

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed with gel electrophoresis unit (Invitrogen Novex, United States) using 12% resolving and 4% stacking gel. 20 µl samples were loaded into each well of the gel and the electrophoresis was then run at 200V for 35 min. The gel was then washed in deionized water, stained in Coomassie brilliant blue and destained until the zones of blue background cleared. Bench mark protein ladder ranging from 10 to 220 kDa was used as the marker.

Effect of pH, Temperature and Storage Stability of Starfruit Protease

Protease was incubated for 24 hrs in different buffers in the pH range of 2.0 to 12.0 before the determination of proteolytic activity. In order to determine the effect of temperature, the protease was incubated at different temperatures in the range of 20 to 80°C at 10°C interval. An aliquot of the protease was analyzed for proteolytic activity. Storage stability of the proteases were determined daily during one week storage at 4°C.

RESULTS AND DISCUSSION

Protein Content, Yield, Protein Concentration and Specific Activity of Starfruit Proteases

Protein content in starfruit was found to be 0.73%, which closely agreed with reports by Ashok *et al.* [15] and USDA [16] where the protein contents are 0.81% and 0.60%, respectively. Proteases purified with acetone resulted in higher yield compared to those purified using 40% ammonium sulfate (Table 1). In addition, yield of proteases from Index 2 (unripe stage) was also higher than those from Index 7 (ripe stage) for both purification methods which is in-line with protein concentration. In Chaurasiya & Hebbar [17] studies, higher protein was obtained in bromelain extracted from partially ripe fruits (12.15 mg/ml) than the fully ripe fruits (11.75 mg/ml). Their results also showed that as ripening progresses bromelain activity decreases. According to Barraclough *et al.* [18], acetone precipitation is more efficient to concentrate the protein. In starfruit protease study, the efficiency of acetone to precipitate the protein can be seen by the higher yield obtained.

Table 1: Yield (%), Protein Concentration (mg/ml) and Specific Activity (CDU/mg) of Purified Starfruit Proteases Prepared Using Different Purification Methods

Purification Method	Maturity Index	Yield (%)	Protein concentration (mg/ml)	Specific activity (CDU/mg)
Acetone	2	1.50	0.054	5407.90
40% ammonium sulfate	2	0.88	0.065	4883.80
Acetone	7	1.20	0.010	26, 667
40% ammonium sulfate	7	0.65	0.018	14,715.74

Molecular Weight Distribution

Figure 1 shows that protein bands with molecular weight range from 10 to 220 kDa are present in Index 2 proteases purified using both acetone and 40% ammonium sulfate. As for Index 7, there is lesser protein bands, which appear extremely, faint than those of Index 2. Between the two purification methods, acetone proteases showed more obvious bands than 40% ammonium sulfate precipitation regardless of maturity stages while protein band at 50 kDa existed in all the proteases. According to Fleischmann *et al.* [19], enzymatically active protein fractions of *Averrhoa carambola* fruit skin consists of four protein bands ranging from 12 to 90 kDa. Wang *et al.* [7] and Siti Balqis & Rosma [20] obtained a single protein band at 60 and 69 kDa, respectively for bitter gourd and *Artocarpus integer* proteases whereas Chaiwut *et al.* [6], obtained protein bands with molecular weights lower than 18.3 kDa for papaya peel extracts and crude latex. In this study, based on the distinct protein band at 50 kDa, most of the proteolytic activity of starfruit proteases might have been contributed by protein characterized by 50 kDa. Presence of very faint bands in the ammonium sulfate proteases indicated that protein denaturation might have occurred during the purification process as proposed by Maldonado *et al.* [21], who claimed that the used of different purification methods might affect the physical or chemical environment of specific proteins differently, thus changing the protein stability or solubility. According to Koay & Gam [22], acetone precipitation is a good method to purify and concentrate protein.

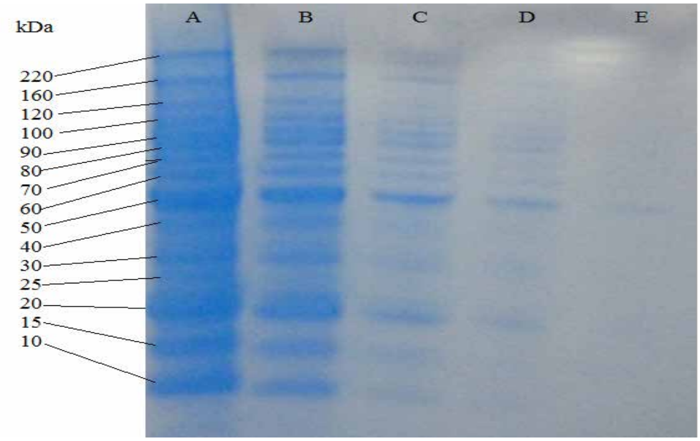

Figure 1: Electrophoretic Profile of Starfruit Proteases at Index 2 and 7 Purified Using Different Purification Methods. A: Protein Marker; B: Index 2 (Acetone Purified Protease); C: Index 2 (40% Ammonium Sulfate Purified Protease); D: Index 7 (Acetone Purified Protease); E: Index 7 (40% Ammonium Sulfate Purified Protease)

Effect of pH on Proteolytic Activity

The proteolytic activity of the Index 2 purified protease was higher compared to the crude extract with the activity range from 100 to 710 CDU for purified protease and 15 to 110 CDU for crude extract while proteolytic activity of acetone purified protease was higher than those purified with 40% ammonium sulfate only up to pH 6 (Figure 2a). Proteolytic activity of crude extract is constant at all pH whereas the purified protease showed maximum activity at pH 8. Higher proteolytic activity of the purified proteases than the crude extract is similarly observed for Index 7 (Figure

2b). In contrary, protease purified with 40% ammonium sulfate was higher than those of acetone up to pH 6. Maximum activity was achieved at pH 8 and 6 for acetone and 40% ammonium sulfate proteases, respectively. pH had no effect on the proteolytic activity of crude extract.

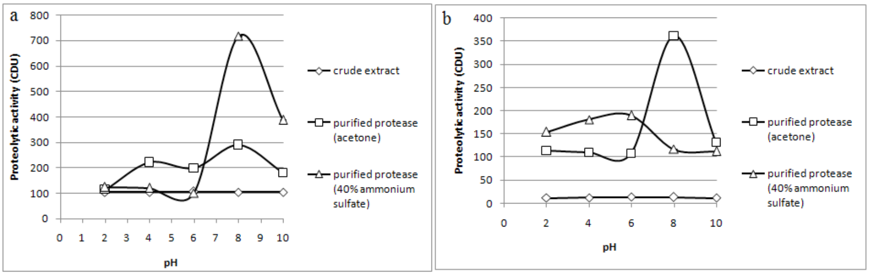

Figure 2: Effect of pH on Proteolytic Activity of Index 2 (a) and Index 7 (b), Starfruit (*Averrhoa carambola L.*) Crude Extract, Protease Purified with Acetone and Protease Purified with Ammonium Sulfate

Maximum bromelain activity of green and ripe pineapple purified using acetone are reported to be at pH 7.5 and 6.5, respectively [23]. Prajapati *et al.* [24] study found that the optimal pH for high enzyme activity for lapsi leaf protease was at pH 7 as compared to pH 8 for the fruit protease. In Bruno's *et al.* [25] study on unripe *Bromelia hieronymi* Mez (Bromeliaceae) purified with cold acetone, the maximum proteolytic activity achieved was at pH 8.5 to 9.5. Corzo *et al.* [26] found that the optimum enzyme activity showed a sharp peak at pH 7.7 for crude bromelain. The use of casein was limited to a neutral to basic pH range because of the low solubility of casein under acidic conditions. Therefore, the optimal pH obtained from an analysis may reflect more on the susceptibility of a substrate to the enzyme at a given pH rather than on the actual activity of the enzyme [20].

Effect of Temperature on Proteolytic Activity

Effect of temperature on protease activity was determined by incubating the starfruit proteases from 20 to 80°C at the interval of 10°C for 15 minutes. Figure 3a shows that the proteolytic activity of the purified proteases are higher than the crude extract while the proteolytic activity of acetone protease is higher compared to those from 40% ammonium sulfate purification method. Maximum activity is achieved at 50°C for acetone protease and 60°C for ammonium sulfate protease. Similar trend is observed for Index 7 except that the proteolytic activity of protease purified with 40% ammonium sulfate is higher than those from acetone purified protease (Figure 3b). Maximum proteolytic activity is achieved at 60°C for 40% ammonium sulfate purified protease and 50°C for acetone protease.

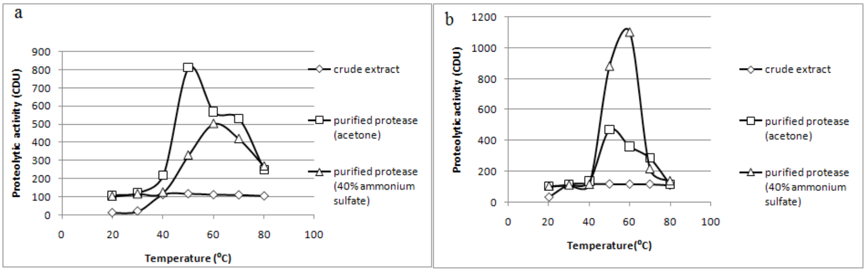

Figure 3: Effect of Temperature on Proteolytic Activity of Index 2 (a) and Index 7 (b), Starfruit (*Averrhoa carambola L.*) Crude Extract, Protease Purified with Acetone and Protease Purified with Ammonium Sulfate

Siti Balqis & Rosma's [20] study on *Artocarpus integer* showed that proteolytic activity steadily increased with the temperature up to 40°C while high activity occurred at 70°C. Priya *et al.* [23] obtained an optimum bromelain activity at 55°C for both unripe and ripe fruit. Corzo *et al.* [26] reported the optimum bromelain activity at 59°C for ripe pineapple. Ibrahim *et al.* [27] study on *Jatropha curcas leaves* protease showed that the optimum temperature for high enzyme activity was at 45°C. Valles *et al.* [28] reported that the optimum bromeliaceae activity was at 60°C for ripe fruits of *Bromelia antiacantha Bertol* while in Asif-Ullah *et al.* [29], the optimum activity for kachri (*Cucumis trigonus* Roxburghi) fruit was at 70°C.

Stability of Proteolytic Activity During Storage

Starfruit proteases of Index 2 maturity stage were incubated for one week at 4°C and the relative proteolytic activity was determined (Figure 4). The proteolytic activity of protease purified with 40% ammonium sulfate rapidly decreased during the first day of storage and slowly decreased thereafter until the relative activity is approximately 57% at day seven. However, for acetone protease, the relative activity remained higher than ammonium sulfate purified protease throughout the seven day storage. Similarly, the activity gradually decreased during storage with relative activity of 70% at day seven. Valles *et al.* [28] reported that the enzyme activity of the protease from ripe fruits of *Bromelia antiacantha* Bertol (Bromeliaceae) was 100% when stored for 180 days at -20°C. Most probably, if starfruit proteases were stored at temperature lower than 4°C a higher relative activity could be restored and the rate of decrease would be lesser.

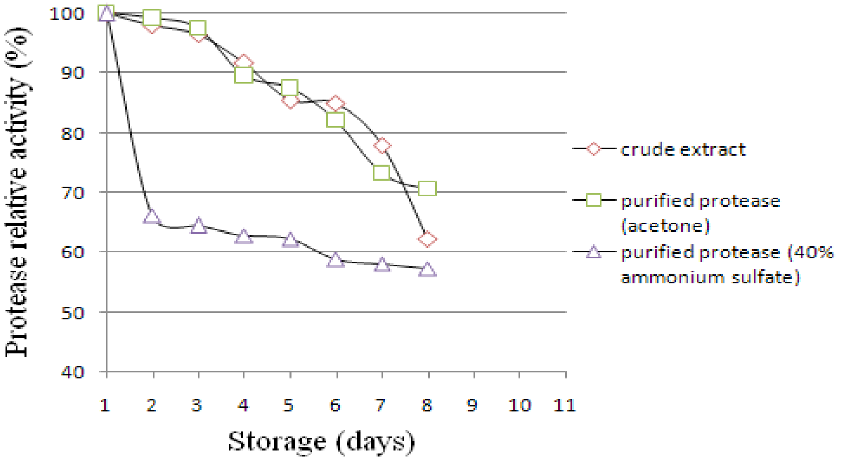

Figure 4: Protease Relative Activity (%) of Index 2 Starfruit (*Averrhoa carambola L.*) Crude Extract, Protease Purified with Acetone and Protease Purified with Ammonium Sulfate During Seven Days Storage at 4°C

CONCLUSION

Protease from starfruit has been successfully extracted and purified by two purification methods comprising of acetone and 40% ammonium sulfate precipitation. Data indicated that the purified protease was stable in the alkaline region with an optimal pH recorded at pH 6 and 8 and at 50 and 60°C. Maturity Index 2 contains higher protein concentration compared to maturity Index 7. The yield for both purification methods ranges from 0.65% to 1.50%. The presence of protein band at 50 kDa indicated that this protein band could have contributed to the proteolytic activity of starfruit protease. The protein bands for protease purified with acetone are more intense compared to those purified with 40% ammonium sulfate. This study suggests acetone precipitation is a better method to purify starfruit protease than ammonium sulfate precipitation. Higher protein concentration and proteolytic activity in Index 2 than 7 suggested that Index 2 starfruit was a better source of protease than Index 7. Therefore, starfruit may serve as another alternative source of plant protease.

REFERENCES

[1] Abdullah, A. G. L., Sulaiman, N. M., Aroua, M. K. & Noor, M.M., 2007. Response surface optimization of conditions for clarification of carambola fruit juice using a commercial enzyme. *Journal of Food Engineering,* 81, pp. 65-71.

[2] Adler-Nissen, J., 1993. Proteases. In *Enzymes in Food Processing* (3rd ed.), Nagodawithana, T., and Reed, G. (eds). Academic Press. San Diego, CA, USA.

[3] Whitehurst, R.J. & van Oort, M., 2010. Enzymes in Food Technology. 2nd eds. Blackwell Publishing, Iowa, USA, pp. 10-15.

[4] Mazumdar, B. C. & Majumder, K. (2003). Methods on physico-chemical analysis of fruits. *Daya Publishing House,* Delhi, India, pp. 156-158.

[5] Azarkan, M., Moussaoui, A. E., Van Wuytswinkel, D., Dehon, G. & Looze, Y., 2003. Fractionation and purification of the enzymes stored in the latex of Carica papaya. *Journal Chromatography,* 790, pp. 38-229.

[6] Chaiwut, P., Pintathong, P. & Rawdkuen, S., 2010. Extraction and three-phase partitioning behavior of proteases from papaya peels. *Process Biochemistry,* 45, pp. 1172-1175.

[7] Wang, L., Wang, M., Li, Q., Cai, T. & Jiang, W., 2008. Partial properties of an aspartic protease in bitter gourd (*Momordica Charantia L.*) fruit and its activation by heating. *Food Chemistry,* 108, pp. 496-502.

[8] Normah, I. & Nur'Ain M.K., 2013. Extraction and partial purification of protease from bilimbi (*Averrhoa bilimbi* L.). *Scientific Research Journal,* 10(2), pp. 1-22.

[9] He, N., Li, Q., Sun, D. & Ling, X., 2008. Isolation, purification and characterization of superoxide dismutase from garlic. *Biochemical Engineering Journal,* 38, pp. 33-38.

[10] Normah, I., Jamilah B., Nazamid, S. & Yaakub, C.M., 1999. A Preliminary Study for the Extraction of Protease from Senduduk (*Melastoma malabatricum)* Leaves. *Proceedings of the 11th National Biotechnology Seminar*, Century Mahkota Hotel, Melaka 22/11-24/11/1999.

[11] AOAC, 2005. Official Method of Analysis, 15th ed. Virginia, Association of Official Analytical Chemistry International.

[12] Chutipongtanate, S., Watcharatanyatip, K., Homvises, T., Jaturongkakul, K. & Thongboonkerd, V., 2012. Systematic comparisons of various spectrophotometric and colorimetric methods to measure concentrations of protein, peptide and amino acid: Detectable limit, linear dynamic ranges, interferences, practicality and unit costs. *Talanta,* 98, pp. 123-129.

[13] Kaneda, M. & Uchikoba, T., 1994. Protease from the sarcocarp of *Trichosanthes bracteata. Phytochemistry,* 35, pp. 583-586.

[14] Ngo, L. T. A., Pham, T. L. & Le, V. V. M., 2008. Purification and endopolygalacturonose from submerged culture of Aspergillus awamori L1 using a two-step procedure: Enzyme precipitation and gel filtration. *International Food Reserach Journal,* 15, pp. 135-140.

[15] Ashok, R., Shoba, H. & Chidanand, D.V., 2011. A study on shelf life extension of carambola fruits. *International Journal of Scientific & Engineering Research,* 9(1), pp. 1-5.

[16] USDA (United States Department of Agriculture), 2003. The USDA national nutrient database for standard reference.

[17] Chaurasiya, R.S. & Hebbar, H.U., 2013. Extraction of bromelain from pineapple core and purification by RME and precipitation methods. *Separation and Purification Technology,* 111, pp. 90-97.

[18] Barraclough, D., Obenland, D., Laing, W. & Carrol, T., 2004. A general method for two dimensional protein electrophoresis of fruit samples. *Postharvest Biology and Technology,* 32, pp. 175-181.

[19] Fleischmann, P., Watanabe, N. & Winterhalter, P., 2003. Enzymatic carotenoid cleavage in starfruit (*Averrhoa carambola*). *Phytochemistry,* 63(2), pp. 131-137.

[20] Siti Balqis, Z. & Rosma, A., 2011. Artocarpus integer leaf protease: Purification and characterisation. *Food Chemistry,* 129, pp. 1523-1529.

[21] Maldonado, A.M., Zomeño, S.E., Baptiste, S.J., Hernández, M. & Jorrín-Novo, J.V., 2008. Evaluation of three different protocols of protein extraction for *Arabidopsis thaliana* leaf proteome analysis by two-dimensional electrophoresis. *Journal of Proteomics,* 71, pp. 461-472.

[22] Koay, S.Y. & Gam, L.H., 2011. Method development for analysis of proteins extracted from the leaves of Orthosiphon *aristatus*. *Journal of Chromatography B,* 879, pp. 2179-2183.

[23] Priya, S.P., Jayakumar, K., Mathai, V. & Babu, S., 2012. Immobilization and kinetic studies of bromelain: A plant cysteine protease from pineapple (*Ananas Comosus*) plant parts. *International Journal of Medical Health Science,* 3(1), pp. 1-7.

[24] Prajapati, S., Sharma, S. & Agrawal, V.P., 2009. Characterization of *Choreospondias axillaris* (Lapsi) fruit protease. *International Journal of Life Science,* 3, pp. 24-31.

[25] Bruno, M.A., Pardo, M.F., Caffini, N.O. & Lopez, L.M., 2003. Heironymain I. A new cysteine peptidase isolated from unripe fruits of *Bromelia hieronymi* Mez (Bromeliaceae). *Journal of Protein Chemistry,* 22(2), pp. 34-127.

[26] Corzo, C. A., Waliszewski, K. N. & Welti-Chanes, J., 2012. Pineapple fruit bromelain affinity to different protein substrates. *Food Chemistry,* 133, pp. 631-635.

[27] Ibrahim, M. A., Olatominwa, J. O., Aliyu, A. B., Bashir, M. & Sallau, A. B., 2012. Partial characterization of protease from the leaves of *Jatropha curcas*. *International Journal of Biology,* 4, pp. 1-7.

[28] Valles, D., Furtado, S. & Cantera, A.M.B., 2007. Characterization of new proteolytic enzymes from ripe fruits of *Bromelia antiacantha* Bertol. (Bromeliaceae). *Enzyme and Microbial Technology,* 3(40), 409-413.

[29] Asif-Ullah, M., Kim, K.S. & Yu, Y.G., 2006. Purification and characterization of a serine protease from *Cucumis trigonus* Roxburghi. *Phytochemistry,* 9(67), pp. 870-875.

Improving Space-Time-Frequency MIMO-OFDM with ICI Self-Cancellation Scheme using Least Square Error Estimator

Nur Farahiah Ibrahim*, Zahari Abu Bakar and Azlina Idris

Faculty of Electrical Engineering,
Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Malaysia
**E-mail: nurfarahiahibrahim@salam.uitm.edu.my*

ABSTRACT

Channel estimation techniques for Multiple-input Multiple-output Orthogonal Frequency Division Multiplexing (MIMO-OFDM) based on comb type pilot arrangement with least-square error (LSE) estimator was investigated with space-time-frequency (STF) diversity implementation. The frequency offset in OFDM effected its performance. This was mitigated with the implementation of the presented inter-carrier interference self-cancellation (ICI-SC) techniques and different space-time subcarrier mapping. STF block coding in the system exploits the spatial, temporal and frequency diversity to improve performance. Estimated channel was fed into a decoder which combined the STF decoding together with the estimated channel coefficients using LSE estimator for equalization. The performance of the system was compared by measuring the symbol error rate with a PSK-16 and PSK-32. The results show that subcarrier mapping together with ICI-SC were able to increase the system performance. Introduction of channel estimation was also able to estimate the channel coefficient at only 5dB difference with a perfectly known channel.

Keywords: *Inter-carrier interference self-cancellation (ICI-SC), Multiple-input Multiple-output (MIMO), Orthogonal Frequency Division Multiplexing (OFDM), channel estimation, Least-square Error (LSE).*

INTRODUCTION

The demand for high speed mobile internet access and high quality streaming multimedia prompted the advancement in digital communication system. This created a demand in the industry for reliable and high capacity link within a limited spectral bandwidth. MIMO-OFDM is a system that uses MIMO antenna configuration with OFDM carrier. It provides the spectral efficiency and multipath fade resistance of OFDM with the throughput and diversity gains of a MIMO system [1-3]. OFDM is an effective method in handling frequency-selective fading by converting a wideband frequency selective channel into parallel narrowband frequency subcarrier.

OFDM has longer symbol duration and higher spectral efficiency, increasing immunity against inter symbol interference (ISI). OFDM orthogonality between subcarriers in time-dispersive environment eliminates crosstalk by the addition of cyclic prefix (CP) [3] . A sufficiently long duration of CP not only prevents ISI but also transforms the linear convolutions onto circular convolutions [4] preserving system's orthogonality and preventing frequency and phase shift errors [5]. MIMO systems offer increase capacity in fading channels with beam forming capabilities, robustness to multi-path delay, spatial diversity and spatial multiplexing [6-9].

Space-time (ST) method improves upon the reliability of data transmission by using multiple transmit antennas. It works by transmitting redundant copies of the data stream to the receiver. Working on the whole block of data at once improves coding and diversity gain and is known as space-time block code (STBC). Space-frequency (SF) coding is applied within a single OFDM block, increasing spatial and frequency diversity gain. When the STBC length is longer than the number of OFDM subcarriers, the code word will span over several OFDM symbols. The whole system is then known as space-time-frequency block code (STFBC) OFDM. Figure 1 shows the STFBC block diagram used in the system. Coding applied across multiple OFDM blocks exploit the spatial, temporal, and frequency diversities available in frequency selective MIMO channels. It was shown in [3,10,11] that coding across blocks in STFBC offers significantly increased diversity order [3,6,9,11,12]. ST performs better in high frequency selective environments since ST requires adjacent OFDM symbols to experience similar fading. ST performance deteriorates if the channel varies quickly against time. SF on the other hand is sensitive to frequency variations. MIMO-OFDM benefits from the implementation of frequency diversity scheme with the use of the orthogonal subcarriers as it removes multipath fading and avoids burst error.

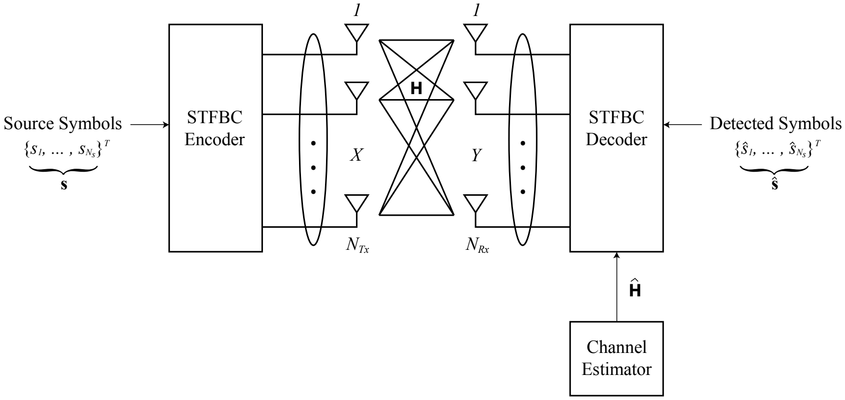
Source Symbols
$\underbrace{\{s_1, \ldots, s_{N_s}\}^T}_{\mathbf{s}}$
STFBC Encoder
1
X
N_{Tx}
H
1
Y
N_{Rx}
STFBC Decoder
Detected Symbols
$\underbrace{\{\hat{s}_1, \ldots, \hat{s}_{N_s}\}^T}_{\hat{\mathbf{s}}}$
$\hat{\mathbf{H}}$
Channel Estimator

Figure 1: Block Diagram of STFBC

Frequency offset in mobile radio channels distort orthogonality between subcarriers resulting in inter carrier interference (ICI). Various methods for reducing ICI exists; frequency-domain equalization, frequency-domain offset estimation/compensation, time-domain equalization, time-domain windowing schemes and ICI-SC. ICI-SC offers the most direct and simplest approach to counter ICI with excellent performance. The only drawback is the reduced channel bandwidth but can be compensated with the use of higher order modulation or higher coding gain.

ICI-SC is also less complex and more efficient compared to other estimation and correction schemes listed [2,13-15]. It is known that the ICI coefficient between two consecutive subcarriers is very small. One of the carriers would carry modulated symbol with a predefined inversed weighting coefficients, “-1” generating a data pair (1, -1) modulated on two adjacent subcarriers (l, l+1). By doing so, generated ICI components can cancel each other at the receiver since ICI in l would be cancelled by the same ICI in l+1. ICI-SC would lower the throughput and bandwidth efficiency by a factor of two due to the redundancy in the carrier. This can be further compensated by using higher order modulation scheme with a high transmission rate [14,16,17].

All reference to ICI-SC scheme showed and proved significant improvement in system performance, ICI and bit-error rate, BER. A data conjugate method which is capable of both error correction and ICI reduction was explained in [2]. The result is compared to data-conversion method and is found to have better performance. Complex conjugate method is shown to have 2dB improvement compared to non-conjugate ICI-SC. A comparison was also made between MIMO (2×2) and SISO with ICI-SC with an improvement of 5dB between them in favor of MIMO [15]. System with no ICI-SC is shown to be adversely affected by carrier frequency offset (CFO).

Channel effects estimates are required at the receiver in order to recover the received data and is used as the channel parameters [5,9]. Two types of channel estimation are pilot based and blind channel estimation. Blind channel estimation uses the correlation between the data being sent and received by the system. A large number of symbols are needed at the receiver in order to extract statistical properties to be used for estimation and usually perform worse than other conventional channel estimation techniques [9]. Blind channel estimation has higher spectral and power efficiency when compared to pilot signaling but is more complex and is only suitable for slow varying channels. Pilot based channel estimation works by obtaining impulse response of pilot symbols inserted in the transmitter.

OFDM due to its orthogonality, was very sensitive to ICI between transmitter and receiver. This may occur due to Doppler shift or multipath channel propagation. The channel is always unknown at the receiver thus channel equalization to compensate for multipath shifts cannot be done. In this paper, a self-cancellation scheme is proposed to combat the effects of ICI. Least-square error (LSE) channel estimator is also implemented to estimate the unknown channel for improved channel equalization in real world application.

SYSTEM INFORMATION

A model of the proposed system is illustrated in Figure 2. Space-time-frequency (STF) implementation consists of two encoders applying both diversities to introduce redundancy in time and frequency domain through multiple transmission antennas. ST signals are further remapped [18] and interleaved to improve immunity to noise. Detail performance and system comparison of different mapping techniques are explained in [1, 9]. ICI-SC is another simple and effective scheme working on the principle of modulating one data symbol on a group of subcarriers with predefined inversed weighting coefficient (X_k, $X_{k=1}$ =–X_k) which will mutually cancel the ICI generated between the two sub-carriers [13].

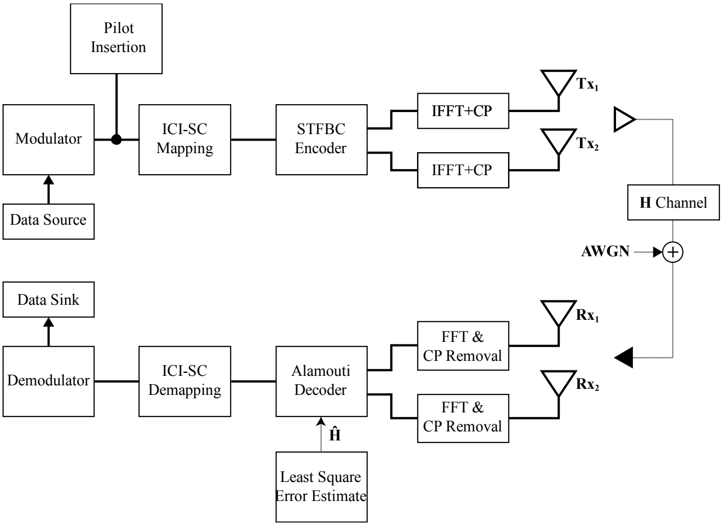

Figure 2: STFBC MIMO-OFDM with CE

The CFO of the transmission link from transmitter antenna m and receiver n antenna $\varepsilon_{m,n}$. is described in Eq.(1). STF mapping is then applied for adjacent, symmetry and median given Eq.(2-4) respectively.

$$Y_n(k) = \sum_{m=1}^{M} C_m(k) H_{m,n}(k) S_{m,n}(0) + I_n(k) + \omega_n(k) \tag{1}$$

$$X'_{2k} = X_k, \quad X'_{2k+1} = -X_k \tag{2}$$

$$X'_k = X_k, \quad X'_{N-1-k} = -X_k \tag{3}$$

$$X'_k = X_k, \quad X'_{N-1-k} = -X_k \tag{4}$$

For ICI-SC scheme, assuming the transmitted symbols are constrained. The modulation is designed to work in such a way that each signal at the k+1th subcarrier (k denotes even number) is multiplied by '-1' and then summed with the one at the kth subcarrier and is represented a Eq.(5) and the ICI coefficient for ICI self-cancellation scheme is denoted in Eq.(6).

$$\begin{aligned} Y''(k) &= Y'(k) - Y'(K+1) \\ &= \sum_{l=0,even}^{N-2} \begin{matrix} X(1)\{S(1-k) - S(l+1-k) \\ \ldots - [S(l-k-1) - S(l-k)]\} + n'(k+1) \end{matrix} \\ &= X(\mathrm{k})\left\{-S(-1) + 2S(0) - S(1)\right\} \\ &\quad + \sum_{l=0,\mathrm{l}\neq \mathrm{k},even}^{N-2} \begin{matrix} X(l)\{-S(l-k-1) - 2S(l-k) \\ \ldots - S(l+1-k)\} + n'(k+1) \end{matrix} \end{aligned} \tag{5}$$

$$S''(l-k) = -S(l-k-1) + 2S(l-k) - S(l+1-k) \tag{6}$$

CHANNEL ESTIMATION (CE)

Block type pilot insertion is suitable for frequency selective channels. Pilot symbols are inserted into each subcarrier with a specific pilot distance. Comb type pilot symbols arrangement is distributed into evenly spaced subcarriers within each OFDM block. It provides higher retransmission rate and better resistance against fast-fading channels [9,19,20].

Different methods exist to estimate the channel based on the received pilot symbols. Minimum mean-square error (MMSE) estimator performs better than a least-square error (LSE) estimator but the computation is significantly more complex and the complexity increases with the number of subcarriers. LSE estimator is widely used due to its simplicity and a good complexity/performance system ratio. However in a more complex system, LSE is often used to get an initial estimates of the pilots which are then further improved via different methods [4,5,7,9,20]. Pilot insertion for channel coefficient estimation, N_p pilot signals are uniformly inserted into X(k) according to Eq.(7), Where L = number of carriers / N_p and $x_p(m)$ is mth pilot carrier value. The frequency response of the channel is defined as in Eq.(8) with $Y_p(k)$ and $X_p(k)$ being the output and the input at the kth pilot sub-carrier respectively.

$$X(k) = X(mL+l) = \begin{cases} x_p(m), & l = 0 \\ \text{inf data} & l = 1,\dots,L-1 \end{cases} \quad (7)$$

$$H_e = \frac{Y_p}{X_p} \qquad k = 0,1,\dots,N_p - 1 \quad (8)$$

Least-square error, LSE estimator method finds the channel estimate $\hat{H}$ in such a way that the following cost function in Eq.(9) is minimized. Setting derivative function with respect to $\hat{H}$ to zero, We have $X^H X\hat{H} = X^H Y$ which gives the solution to the LSE channel estimation as in Eq.(10).

$$\begin{aligned} J = \left(\hat{H}\right) &= \left\|Y - X\hat{H}\right\|^2 \\ &= \left(Y - X\hat{H}\right)\left(Y - X\hat{H}\right)^H \\ &= Y^H Y - Y^H X\hat{H} - \hat{H}^H X^H Y - \hat{H}^H X^H X\hat{H} \end{aligned} \quad (9)$$

$$\hat{H}_{LS} = (X^H X)^{-1} X^H Y = X^{-1} Y \quad (10)$$

Let us denote each component of the LSE channel estimate $\hat{H}$LS by $\hat{H}$LS(k), k = 0,1,2,…,N-1. In the decoder, taking the received signal in k and k+1 time slot as in the following and assuming channel remains constant in both frame, we get the following matrix in Eq.(11) where Y_n^m is the received information, Z_n^m is the noise and Hmn is the channel from nth received antenna to the mth transmit antenna. X_1, X_2 are the transmitted symbols. Combining equations at time k and k+1 gives Eq.(12).

$$\begin{bmatrix} Y_1^1 \\ Y_2^1 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} Z_1^1 \\ Z_2^1 \end{bmatrix}$$

$$\begin{bmatrix} Y_1^2 \\ Y_2^2 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} -X_2^* \\ X_1^* \end{bmatrix} + \begin{bmatrix} Z_1^2 \\ Z_2^2 \end{bmatrix} \quad (11)$$

$$\begin{bmatrix} Y_1^1 \\ Y_2^1 \\ Y_1^{2*} \\ Y_2^{2*} \end{bmatrix} = \begin{bmatrix} \mathrm{H}_{11} & \mathrm{H}_{12} \\ \mathrm{H}_{21} & \mathrm{H}_{22} \\ \mathrm{H}_{12}^{*} & -\mathrm{H}_{11}^{*} \\ \mathrm{H}_{22}^{*} & -\mathrm{H}_{21}^{*} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} Z_1^1 \\ Z_2^1 \\ Z_1^{2*} \\ Z_2^{2*} \end{bmatrix} \qquad (12)$$

Thus to solve for X1 and X2, we need to find the inverse of the channel estimation matrix $\hat{\mathrm{H}}$ given by Eq (10). The pseudo inverse of the equation results in the decoding of the transmitted symbols.

$$\begin{bmatrix} \\ \end{bmatrix} = \left(\mathrm{H}^{\mathrm{H}}\mathrm{H}\right) \quad \mathrm{H}^{\mathrm{H}} \begin{bmatrix} \\ \\ {}^{2*} \\ {}^{2*} \end{bmatrix} \qquad (13)$$

SYSTEM PERFORMANCE EVALUATION

The symbol-error rate, SER curves are used to evaluate and compare various aspects of the proposed system. System is designed and simulated in Matlab®. A quasi-static Rayleigh fading channel with varied additive white Gaussian noise, AWGN is used throughout the simulations. Rayleigh fading is selected since it closely model the statistical effect model of tropospheric and ionospheric radio signal propagation and urban city environment with no line of sight, LOS.

Figure 3 illustrates the performance difference between different mapping techniques. Selected ST median subcarrier mapping is able to provide an average improvement of 3.5dB compared to unmapped carriers. This shows that interleaving and remapping subcarriers would increase immunity against burst noise and long channel fade. Median subcarrier mapping will be used in the paper due to better performance compared to the other mapping techniques. Implementation of ICI-SC in the system was able to give a significant improvement of 4dB as shown in Figure 4. The shift in carrier frequency due to the applied Doppler in the transmission medium was mitigated effectively by ICI-SC.

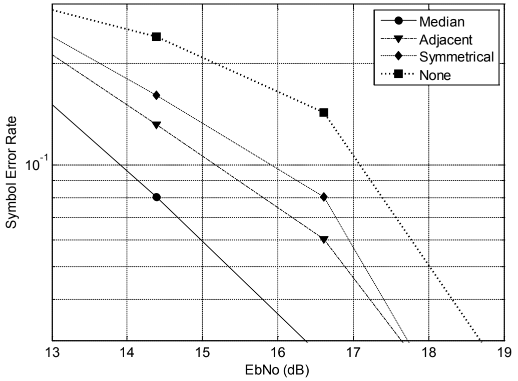

Figure 3: Performance Comparisons of Different ICI-SC Mapping Techniques

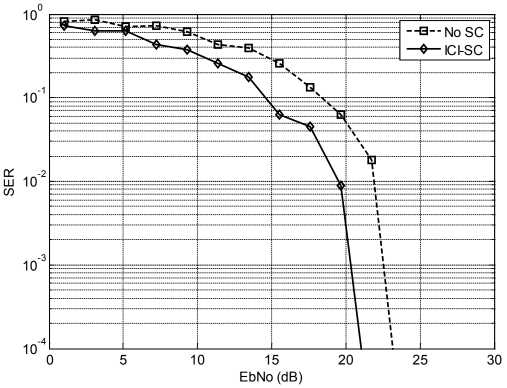

Figure 4: Performance Improvement with and without ICI-SC Using Median Method

A comparison of the LSE estimator is done between known channel and estimated channel together with the result without channel estimation. Rayleigh fading is used as the propagation channel. PSK-16 and PSK-32 modulations were performed to examine how the system performance scales with larger symbol size. Pilot symbols are arranged in block with 16 subcarriers acting as pilots with a spacing of eight subcarriers in between. Spectral energy of pilots is amplified by a factor of two. The increase in pilot symbols and reduction of pilot distance between each other would increase the overall system performance at the expense of actual data transmission.

Comb pilot arrangement is found to be dependent on the rigorousness of the pilot symbol placement and the estimator performance is closely related to the estimator algorithm implemented. The final performance of the proposed system is evaluated with the parameters in Table 1. The performance of the system was compared with two symbol sizes in a multi-path Rayleigh fading channel with Doppler shift. The performance difference is linearly static between both PSK-16 and a PSK-32 modulation. The LSE comb pilot symbol estimation was able to estimate channel $\hat{H}$ providing a performance difference of only 5dB compared to the perfect known channel H as shown in Figure 5. A blind channel (no estimation) provided no improvement whatsoever. The whole data block is irrecoverable.

Table 1: General Simulation Parameters for STFBC MIMO-OFDM with ICI-SC with LSE Estimator

Parameters	Value
Number of Subcarriers	128
Size of Subcarrier	8192
Cyclic Prefix	256
Modulation	PSK-16 & 32
Pilot Arrangement	Block
Number of Pilot SC	16
Pilot SC Spacing	8

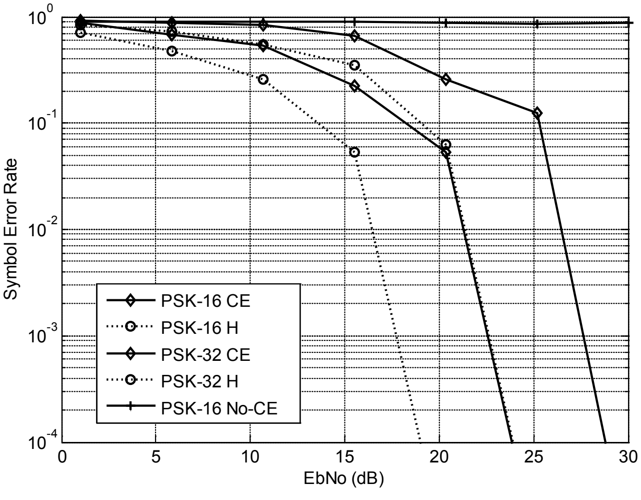

Figure 5: Performance Comparison of Least Square Error Estimation with Different Modulation Techniques

CONCLUSION

This paper explored different channel coding approach towards STF diversity to achieve maximum diversity order. MIMO system in OFDM proved to be versatile in combating ISI effects with its orthogonal properties. However, CFO significantly impact OFDM performance. ICI-SC has shown great performance with significant improvements over its simplicity to combat CFO in OFDM. The subcarrier mapping methods further improved upon STF bare performance. This approach offers the benefits of simplicity by eliminating the interpolator and combining the channel equalization in the STF MIMO decoder. Applying the channel estimator to recover the channel coefficient H gave significant improvement over a bare system with no channel estimation. This shows the importance of channel estimation in an unknown channel medium. Further improvements to the estimated $\hat{H}$ channel could be implemented with more complex and better performing minimum mean-square error (MMSE) and singular value decomposition (SVD) paired with channel interpolation as presented in references [4,5,19].

REFERENCES

[1] A. Idris, K. Dimyati, S. K. S. Yusof, and D. M. Ali, 2011. Pairwise Error Probability of a New Subcarrier Mapping Scheme (ICI-SC Technique) for STFBC MIMO-OFDM System, *Australian Journal of Basic and Applied Science, vol. 5.*

[2] A. Idris, K. Dimyati, and S. K. S. Yusof, 2008. Interference Self-Cancellation Schemes for Space Time Frequency Block Codes MIMO-OFDM System, *IJCSNS Internatonal Journal of Computer Science and Network Security, vol. 8*, pp. 139-148.

[3] A. F. Molisch, M. Z. Win, and J. H. Winters, 2002. Space-Time-Frequency (STF) Coding for MIMO-OFDM Systems, *IEEE Communications Letters, vol. 6,* pp. 370-372.

[4] A. U. Ahmed, S. C. Thompson, and J. R. Zeidler, 2008. Channel Estimation and Equalization for CE-OFDM in Multipath Fading Channels.

[5] J. J. V. d. Beek, O. Edfors, M. Sandell, S. K. Wilson, and P. O. Borjesson, 1995. On Channel Estimation in OFDM Systems, *Vehicular Technology Conference, vol. 2.*

[6] L. M. Cortes-Pena, 2009. MIMO Space-Time Block Coding (STBC): Simulations and Results, Personal and Mobile Communications.

[7] M. E. S. Coleri, A. Puri & A. Bahai. A Study of Channel Estimation in OFDM Systems, IEEE Vehicular Technology Conference.

[8] S. Adegbite, B. G. Stewart, and S. G. McMeekin, 2013. Least Square Interpolation Methods for LTE System Channel Estimation over Extended ITU Channels, *International Journal of Information and Electronics Engineering, vol. 3.*

[9] Y. S. Cho, J. Kum, W. Y. Yang, and C.-G. Kang, 2010. MIMO-OFDM Wireless Communications with MATLAB: Wiley.

[10] Z. S. a. K. J. R. L. W.Su, 2005. Towards Maximum Achieveable Diversity in Space, Time and Frequency : Performance Analysis and Code Design, *IEEE Transactions on Wireless Communications, vol. 40.*

[11] M. Qinghua, Y. Luxi, and H. Zhenya, 2008. Diversitry Analysis of Space-Time-Frequency Coded Broadband MIMO-OFDM System with Correlation Across Space Time and Frequency, *Frontier in Electrical and Electronic Engineering, vol. 3,* pp. 295-300.

[12] H. J. A. R. C. V. Tarokh, 1999. Space-Time Block Codes from Orthogonal Designs, *IEEE Transactions on Information Theory, vol. 45.*

[13] B.S. Kumar, K.R.S. Kumar, and R. Radhakrishnan, 2009. An Efficient Inter Carrier Interference Cancellation Schemes for OFDM System, 2009. *International Journal of Computer Science and Information Security, vol. 6.*

[14] A. Idris, K. Dimyati, and S. K. S. Yusof, 2011. Evaluating A New Subcarrier Mapping ICI-SC Scheme Using Linear Maximum Likelihood Alamouti Combiner (LMLAC) Decoding Technique, *Journal of Engineering Science and Technology, vol. 6,* pp. 664-673.

[15] N. Fisal, S. Kamilah, and A.Izzati. Intercarrier Interference Self-Cancellation for Space-Time-Frequency MIMO-OFDM System.

[16] L. Zhao, 2006. A New ICI Self-Cancellation Scheme Based on Repeated Symbol in OFDM Systems, 2006. International Conference on Communications, Circuits and Systems Proceedings, vol. 2, pp. 1216-1220.

[17] Y.H. Peng, Y.C. Kuo, G.R. Lee, and J.H. Wen, 2007. Performance Analysis of a New ICI-SELF-Cancellation-Scheme in OFDM Systems, IEEE, pp. 1333-1338.

[18] A. Idris, K. Dimyati, S. K. S. Yusof, D. M. Ali, and N. Ya'acob, 2011. CIR and BER Performance of STFBC in MIMO OFDM System, *Australian Journal of Basic and Applied Science, vol. 5,* pp. 3179-3187.

[19] M. H. Hsieh and C. H. Wei, 1998. Channel Estimation for OFDM Systems based on COMB-Type Pilot Arrangement in Frequency Selective Fading Channels, *IEEE Transactions on Consumer Electronics, vol. 44.*

[20] F. D. Y. Sun, 2009. Pilot Aided Channel Estimation for MIMO-OFDM Systems, London Communication Symposium.

Research Trends in Microarray Data Analysis: Modelling Gene Regulatory Network by Integrating Transcription Factors Data

Farzana Kabir Ahmad* and Siti Sakira Kamaruddin

Computational Intelligence Research Cluster,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
**E-mail: farzana58@uum.edu.my*

ABSTRACT

The invention of microarray technology has enabled expression levels of thousands of genes to be monitored at once. This modernized approach has created large amount of data to be examined. Recently, gene regulatory network has been an interesting topic and generated impressive research goals in computational biology. Better understanding of the genetic regulatory processes would bring significant implications in the biomedical fields and many other pharmaceutical industries. As a result, various mathematical and computational methods have been used to model gene regulatory network from microarray data. Amongst those methods, the Bayesian network model attracts the most attention and has become the prominent technique since it can capture nonlinear and stochastic relationships between variables. However, structure learning of this model is NP-hard and computationally complex as the number of potential edges increase drastically with the number of genes. In addition, most of the studies only focused on the predicted results while neglecting the fact that microarray data is a fragmented information on the whole biological process. Hence, this study proposed a network-based inference model that combined biological knowledge in order to verify the constructed gene regulatory relationships. The gene regulatory network is constructed using Bayesian network based on low-order conditional independence approach. This technique aims to identify from the data the dependencies to construct the network structure, while addressing the structure learning problem. In addition, three main toolkits such as Ensembl, TFSearch and TRANSFAC have been used to determine the false positive edges and verify reliability of regulatory relationships. The experimental results show that by integrating biological knowledge it could enhance the precision results and reduce the number of false positive edges in the trained gene regulatory network.

***Keywords**: Gene regulatory network, Bayesian Network, Heterogeneous data, Transcription Factors*

INTRODUCTION

The invention of microarray technology can be considered as the latest technological breakthrough in molecular biology. Microarray experiments allow expression levels of thousands of genes to be monitored at once to provide complete transcription information in the cells. This revolutionized approach has provided a large amount of data from which a lot of knowledge can be explored. Despite the achievement of microarray technology that constantly improves laboratory methods and is prominently being used in biological researches, the major advances of the field is actually derived from the enhanced analysis methods. Due to its high-throughput nature, microarray data usually pose several challenges in terms of data analysis. Thus, computational approaches are generally necessary to divulge the molecular mechanism of cancerous cell and gain holistic view of how all these genes interact. Microarray data analysis generally consists of two major parts, namely the initial stage and the exploratory data analysis stage. The initial stage function is to prepare the raw data for rigorous analyses, as well as playing a crucial role to avoid any key factor that may affect subsequent results. On the other hand, the exploratory data analysis is an approach used to examine microarray data for the purpose of answering research questions.

In conjunction with this invention, network-based classification approach has been used in identifying gene markers that present the maximum discrimination power between cancerous and normal cells. Hence, identifying gene regulatory network has been an impressive research goal and new trend in computational biology. This interesting topic has become one of the vital research areas in microarray data analysis. Gene regulatory network (GRN) is a set of molecular components that includes genes, proteins and other molecules, which collectively accomplish cellular functions as these molecules interact with each other [1]. The fundamental idea behind GRN analysis is to discover regulator genes by examining gene expression patterns. Notably, some genes regulate other genes, which mean that the amount of a gene expressed at a certain time could activate or inhibit the expression of another gene. Thus, changes in the expression levels of particular genes across a whole process, such as response to certain treatments would provide information that allows reconstruction of GRN using reserve engineering technique. Such data-driven regulatory networks analysis ultimately offers clearer understanding of the genetic regulatory processes, which are normally complex and intricate. Furthermore, it would bring significant implications in the biomedical fields and many other pharmaceutical industries.

Numerous studies have reported that GRN can possibly assist researchers in suggesting and evaluating innovative hypotheses in the context of genetic regulatory processes [2-3]. Various mathematical and computational methods have been used to model GRN from microarray data, including Boolean network, pair-wise comparison, differential equations estimation, Bayesian network and other techniques. Amongst these, the Bayesian network model attracts the most attention and has become the prominent technique because it can capture linear, nonlinear, combinatorial, stochastic and casual relationships between variables. Compared to other methods, Bayesian network model establishes considerable relationships between all genes in the system. Thus, Bayesian network is used in this study to analyze gene regulatory processes and to model gene relationships for breast cancer metastasis. However, the structure learning of Bayesian network is NP hard.

Furthermore, most of computational methods that were used in classification approach mainly focused on prediction and/or performance results while neglecting the interaction among genes that determine the disease phenotype. Additionally, microarray data only provides fragmented information regarding the whole biological processes. As a result, combining microarray with other biological knowledge is important in order to attain better understanding of cancer-related process and improve predictive power of inference model. Based on above trends and gaps, this study has developed a network-based inference model from gene expression data. Bayesian network has been recognized as an outstanding method to model GRN.

However, structure learning of this model is NP-hard and computationally complex as the number of potential edges increase drastically with the number of genes. Thus, low-order conditional independence method has been proposed to cater the high-throughput data. Although, the proposed method has significantly outperformed compare to other methods and had achieved better performance, the inferred network may still be deficient in terms of biological knowledge. Furthermore, reliability issues and false positive edges are other problems. For these reasons, data integration such as transcription factors has appeared as a right way to revise and check on regulatory relationships. In addition, biological knowledge is combined to further improve the proposed network.

The remainder of this paper is organized as follows. Section 2 describes some previous works, in which different kinds of biological data were utilised to achieve better construction of GRN. Section 3 on the other hand, presents the proposed method. Section 4 meanwhile presents experimental results and discussion. Finally, Section 5 offers concluding and future direction remarks.

RELATED WORKS

In recent years, the increasing amount of genomic data such as gene expression and proteomic with publicly available databases is another trigger that has opened a new promising approach to combine various data types. Although these data varies in term of size, formats and types, it provides different, partly independent and complementary information on the whole genome. In principle, if cellular biology knowledge is complete, one can infer the genomic interactions given the activity of each molecule at a time. Unfortunately, such network is not yet available for any cell type [4]. Thus, the best option is to integrate diverse biological data that presents fragmented information and seek a better explanation for the development at a system level.

Aligned with these motivations, several studies have combined different types of data to obtain comprehensive network [5-6]. Mainly, there are two different categories to combine data: (1) homologous data integration and (2) heterogeneous data integration. Homologous data integration is defined as the use of similar data type (for example combination of multiple microarray datasets from different studies). Meanwhile, the latter categories integrate different data types across or within studies to seek for better clarify of information provided by a single data type.

The main idea for homologous data integration is to increase the number of samples to address the issue of high dimensional data. Most studies in homologous data integration have focused on comparing two or more related datasets to identify significant genes that can distinguish different group of samples (e.g. disease and normal samples). For example, [7] have combined multiple microarray datasets to classify common transcription profiles that are universally activated in most cancer types. Generally two main methods have been used in combining homologous dataset namely a) meta-analysis method and b) effect size method.

Unlike homologous data integration, where it used similar data types, heterogeneous data integration mainly focuses on applying various data sources to ensure the reliability of results obtained. Among the popular data integration is gene expression and proteomic data. Protein is the end product of translation process and is also used as a trigger to initiate the expression of other genes. Therefore, the combination of these data type is reasonable to most researchers. Besides that, large number of researchers also utilized transcription factor binding sites (TFBS) to verify the GRN. Like protein, TFBS is another complementary data to measure cellular state. Hence, more recent works have explored data integration of external knowledge to identify transcription factors and their target genes [5-6]. Transcription factors are very essential in regulating gene expression. Motivated by this fundamental concept, transcription factors have been used in this research to discover significant biological information from high-throughput data.

METHODS

This section describes the Bayesian network with the low-order conditional independence along with integration of transcription factor in examining gene regulatory processes for breast cancer metastasis.

Bayesian Network

The Bayesian network is a graphical model that was introduced by Pearl and Wright in 1980s [8]. To deal with a large number of genes in microarray data, this research defines the Bayesian network, *BN* as: $BN = (G,P)$ where $G= (X, E(G))$ is a DAG with a set of variables X representing $\{X_i; i \in V\}$, and $E(G) \subseteq X_i * X_j$ (set of pairs that represents the dependent among *v* variables). The element *E* is an edge from node X_i to X_j, indicating X_i is a parent to X_j. On the other hand, *P* corresponds to joint distribution on the variables in the network. The $Pa(V)$ represents the parent for a set of vertex *V* and can be defined as:

$$Pa(X_i, G) = \{X_j, \text{ such that } (X_i, X_j) \in E(G); j \in V\} \tag{1}$$

where $Pa(X_i, G)$ is the parent of X_i in the graph, *G* and having nodeX_j pointing toward X_i.

Bayesian structure learning is a NP hard problem [9]. As the number of possible structures in a Bayesian network grows exponentially with respect to the number of variables (large number of genes in microarray dataset), exhaustive search of all possible structures becomes computationally expensive. Thus, structure learning of Bayesian network is currently a challenging task in modelling GRN. In this study we proposed to low-order conditional independence and its variants, full-order conditional independence, to construct a GRN. For additional technical details on this proposed method please refer to Ahmad, Deris and Othman [10].

Integration of Transcription Factors in GRN

The Bayesian network has emerged as a powerful tool to infer gene regulatory process. However, this method is usually confronted with structure learning problems in handling large-scale of gene expression data. To address such problem, Bayesian network with constraint-based algorithm is proposed as explained in Section 3.1. The basic idea is to develop GRN by measuring the dependencies among nodes of the given data. Low-order conditional independence is used to examine the relationships between genes. Although the proposed method has increased the accuracy of inferred network, such gene network is solely based on the microarray data and is often insufficient for rigorous analysis. In many cases, microarray data is often daunted by noisy, incomplete data and misleading outliers, which can produce high number of false positive edges. Accordingly, an inferred GRN may contain some incorrect gene regulations that are unreliable from the biological point of view. Thus, integration of biological knowledge into gene network has become necessary to overcome the problem.

This study used transcription factors concept in determining the gene regulatory relationship. Three main bioinformatics toolkits have been used in extracting transcription factor and region binding, which includes (1) Ensembl, (2) TFSearch and (3) TRANSFAC. Activation or inhibition for each regulatory relationship is determined by the following definition:

1. Activation $\left(X \xrightarrow{+} Y\right)$ – IF X is over-expressed (X = positive), THEN Y is over-expressed (Y = positive), IF X is under-expressed (X = negative), THEN Y is under-expressed (Y = negative).

2. Inhibition $\left(X \xrightarrow{-} Y\right)$ – IF X is over-expressed (X = positive), THEN Y is under-expressed (Y = negative), IF X is under-expressed (X = negative), THEN Y is over-expressed (Y = positive).

Dataset Description

We tested this proposed method using a data set of 97 breast cancer microarray from van‘t Veer *et al.* [11]. These cohorts of breast cancer patients are 55 years old or younger. We obtained this data from the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA, 2006). Among the remaining 97 samples, 46 developed distant metastasis within 5 years and 51 remained metastasis free for at least 5 years. DNA microarray analysis was used by van‘t Veer to determine the expression levels of approximately 25,000 genes for each patient.

EXPERIMENTAL RESULTS AND DISCUSSIONS

To obtain insights into the mechanism of gene regulation and how gene mutations act to turn on tumour development and metastasis progression in a cellular network context, the proposed method is executed on the breast cancer dataset producing a GRN as shown in Figure 1.

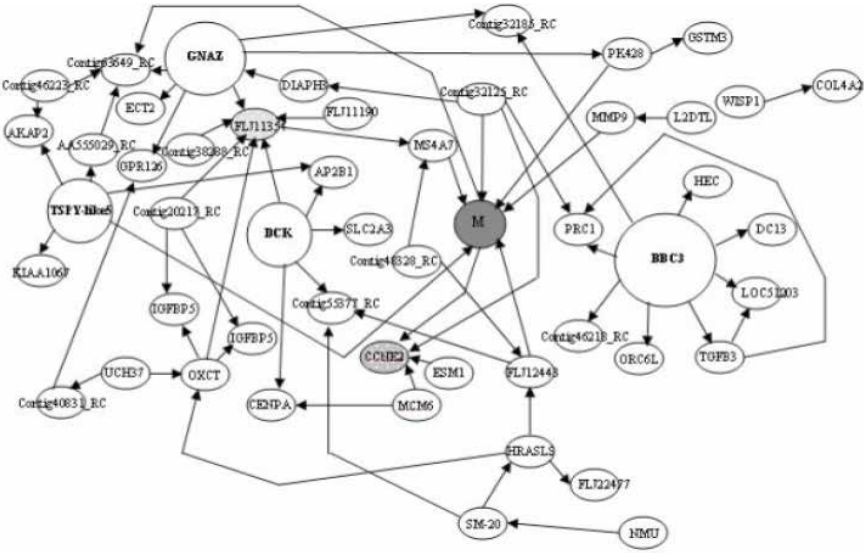

Figure 1: The GRN for Breast Cancer Metastasis Using the Low-Order Conditional Independence Method

This learned network revealed a group of genes which are primarily associated with causing metastasis, M. The larger nodes in the graph specify the genes when expressed at different levels lead to a major effect on the status of other genes (e.g., on or off). Meanwhile, the light-shaded nodes denote the highly regulated genes. Four genes that are found to regulate the expression levels of other genes are: BBC3, GNAZ, TSPY-like5 (TSPY5), and DCK. Two genes are highly regulated: FLJ11354 and CCNE2. This GRN involved 50 genes associated with metastasis, M, and 39 of them are annotated.

To verify the regulatory relationship, this study has tested each relationship using biological integrated data. Three main toolkits such as

Ensembl, TFSearch and TRANSFAC have been used to determine the false positive edges. The experimental results have shown (Table 1) that by integrating heterogeneous data from these sources, the number of false positive edges can be reduced. Accordingly, 258 interactions are found to be biologically related. These interactions have fulfilled the biological test and hypothesis that are set earlier. The results show that the proposed method works better with biological knowledge processing in comparison to network that rely on microarray only since the number of FP edges are discovered decline and the precision result has increased by 2.48% .

Table 1: Precision Results for Cellular Network Without/with Biological Knowledge Processing. Both Networks are Constructed with 5000 genes (TP = True positive; FP = False positive)

Method	Total Edges	FR Edges	TP Edges	Precision %
Low-order conditional independence without biological knowledge	303	58	245	80.85
Low-order conditional independence with biological knowledge	258	43	215	83.33

Table 2 on the other hand illustrates the percentage of activation/ inhibition for each regulatory relationship. Based on the results that have been obtained, this study has discovered that there are five main gene regulators namely BBC3, TGFB3, L2DTL, GNAZ, and TSPY-like 5 that could possibly regulate the expression levels of other genes and highly correlated with breast cancer metastasis. BBC3, TGFB3 and L2DTL are gene regulators that activate the regulation of other genes. These genes mainly inhibit DNA synthesis and induce apoptosis which is the process of programmed cell death (PCD) that causing cells death or damaged. Meanwhile, GNAZ, and TSPY-like 5 are another two genes that most likely activate the expression of other genes which could eventually obstruct the signal transduction pathway.

Table 2: The Percentage of Activation and Inhibition

Regulatory Relationship	Activition %	Inhibition %
BBC3 → HEC	34.38	65.62
BBC3 → DC13	32.81	67.19
BBC3 → PRC1	34.38	65.62
BBC3 → ORC6L	32.81	67.19
BBC3 → Contig46218_RC	35.94	64.06
BBC3 → LOC51203	35.94	64.06
BBC3 → TGFB3	65.63	34.37
BBC3 → Contig32185_RC	31.25	68.75
GNAZ → PK428	51.56	48.44
GNAZ → Contig32185_RC	65.63	34.37
GNAZ → ECT2	54.69	45.31
GNAZ → Contig63649_RC	64.06	35.94
GNAZ → FLJ11354	53.13	46.87
GNAZ → GPR126	60.93	39.07
TSPY-like 5 → AP2B1	65.63	34.37
TGFB3 → PRC1	21.57	78.13
TGFB3 → LOC51203	20.31	79.69
L2DTL → MMP9	29.69	70.31

CONCLUSION AND FUTURE WORKS

This paper describes the need to integrate diverse data integration for better interpretation of GRN model. Two types of data integration approaches have been comprehensively explained; (1) homologous data integration and (2) heterogeneous data integration. Since most GRN models are mainly implemented based on microarray data, issues like reliability and quality concern are also debated by many researchers. The best available alternative is to integrate different data to address this problem and obtain a better understanding of the underlying gene regulatory mechanisms. Furthermore, with the currently available and enormous public databases, this effort appears to be the most promising since it utilizes the independent and complementary information to answer research questions. The use of transcription factors to identify relevant regulatory interactions is the key idea in this research. In achieving this, three main bioinformatics toolkits for instance Ensembl, TFSearch and TRANFAC have been used. Each of these tools is used to apprehend the concept of biological intrinsic features of transcription factor and promoter. Based on the experiments that were conducted, 258 out of 303 interactions are identified to be biologically relevant. Furthermore, this study has discovered that there are five main gene regulators namely BBC3, TGFB3, L2DTL, GNAZ, and TSPY-like 5 that play essential role in breast cancer metastasis. In the future, many more different data types will be integrated to obtain more insightful view of GRN and further facilitate our understanding of cancer growth.

ACKNOWLEDGEMENT

The authors sincerely thank the Research Management Centre, Universiti Utara Malaysia, for the financial support and facilities provided to complete this study.

REFERENCES

[1] F. Yavari, F. Towhidkhah and S. Gharibzadeh, 2008. Gene regulatory network modeling using Bayesian networks and cross correlation, *Biomedical Engineering Conference (CIBEC).*

[2] Z. Huang, J. Li, H. Su, G. Watts and H. Chen, 2007. Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining, *Decision Support Systems, vol. 43,* no. 4, pp. 1207-1225.

[3] W. Zhao, E. Serpedin and E. Dougherty, 2008. Recovering Genetic Regulatory Networks from Chromatin Immunoprecipitation and Steady-State Microarray Data', *EURASIP Journal on Bioinformatics and Systems Biology, vol. 2008,* pp. 1-12.

[4] I. Ulitsky, 2009. Network-based algorithms for analysis of heterogeneous biomedical data, Ph.D., Tel-Aviv University, Israel.

[5] C. Huttenhower, K. Mutungu, N. Indik, W. Yang, M. Schroeder, J. Forman, O. Troyanskaya and H. Coller, 2009. Detailing regulatory networks through large scale data integration, *Bioinformatics, vol. 25, no. 24,* pp. 3267-3274.

[6] C. Kaleta, A. Göhler, S. Schuster, K. Jahreis, R. Guthke and S. Nikolajewa, 2010. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis, *BMC Syst Biol*, *vol. 4, no. 1,* p. 116, 2010.

[7] D. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. Chinnaiyan, 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, *Proceedings of the National Academy of Sciences, vol. 101, no. 25,* pp. 9309-9314.

[8] J. Pearl, 1998. *Probabilistic reasoning in intelligent systems*. San Fransisco, Cal.: Morgan Kaufmann.

[9] D. Chickering, D. Heckerman and C. Meek, 2004. Large-sample learning of Bayesian Networks is NP-hard, *Journal of Machine Learning Research, vol. 5,* pp. 1287-1330.

[10] F.K. Ahmad, S. Deris and N.H. Othman, 2012. The inference of breast cancer metastasis through gene regulatory networks, *Journal of Biomedical Informatics*, *vol. 45, no. 2,* pp. 350-362.

[11] L. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart and Mao, 2002. Gene expression profiling predicts clinical outcome of breast cancer, *Nature, vol. 415,* pp. 530-536.

GUIDELINES FOR SUBMISSION OF ARTICLES

Emphasis is place on direct and clearly understood communication, originality and scholarly merit. Only original manuscripts will be accepted and copyright of published papers will be vested in the publisher.

Submitted manuscripts should be written in English with production - quality figures. Manuscripts should be typed using Times New Roman (point 12), 1.5 spacing and wide margins (2.54 cm left, right, top and bottom). The manuscript should include the title of the paper; the name, address, contact numbers and email of the correspondence author; a short abstract of between 200 to 300 words, which clearly summarizes the paper with 4-5 keywords. Submission should be limited to a maximum of 20 typed pages.

Tables should be included within the text where appropriate and must be numbered consecutively with Arabic numerals and have titles that precede the table. Similarly, all figures must be numbered and a detailed caption should be provided below each figure. Figures should be embedded within the text where appropriate. Glossy photographs when required should be scanned to a suitable resolution (1200 dpi), which enables quality reproduction.

References should be numbered in ascending order and cited within square brackets, e.g. [1], [3-5], in the main body of the text. References included at the end of the manuscript are to be in the order that they appear in the main body of text. Some entry samples are as follows:

1. A. B. Author, 2000. Title of Book, ABC Press, Kuala Lumpur, Malaysia.

2. C. D. Author and E. F. Author, 1999. Title of Paper, Journal Name, vol. 10, pp. 32-45.

3. G. H. Author, 2006. Title of the conference paper, in Proceedings of the 2000 IEEE International Symposium on Circuits and Systems, Geneva, Switzerland, pp. 100-105.

Manuscripts submitted to the journal will be initially screened by the Chief Editor, to determine appropriateness. Only those manuscripts considered of a sufficiently high standard will proceed to undergo a double blind review.

Upon completion, the manuscripts will be passed to an editorial board member for appraisal of their value. Additionally, they will be reviewed by an expert in that discipline.

For further information, kindly e-mail: SRJ_Editor@salam.uitm.edu.my.

MS ISO/IEC 17021:2006
QS 12022007 CB 06

ISO 9001:2008 Certificate No. KLR 0404089

ISSN 1675-7009

9 771823 779008

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA