Лейцин-Баялаг Давталтат Уургийн Давталтыг Регуляр Илэрхийллээр Таних

Г.Ундрал*, П.Энхбаяр**

* Шинжлэх Ухааны Академи; Физик, Технологийн Хүрээлэн; Симуляци, Параллель Тооцооллын Лаборатори

** Монгол Улсын Их Сургууль; Хэрэглээний Шинжлэх Ухаан, Инженерчлэлийн Сургууль; Мэдээлэл, Компьютерийн Ухааны Тэнхим

*undak.1728@yahoo.com, **enkhbayar.p@seas.num.edu.mn

Хураангуй

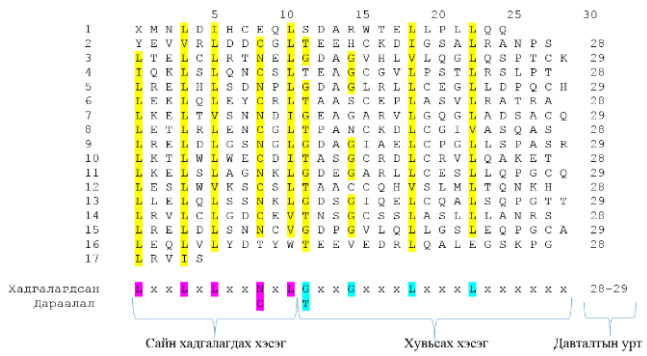
Лейцин-Баялаг Давталт агуулсан дараалал нь вирус, бактер, архей, эукариотоос ялгасан 20000 гаруй уураг агуулагддаг. Лейцин-баялаг давталтат уургийн бүтэц нь морины тах, эллипс эсвэл суперхеликс хэлбэртэй байх ба дотоод хүнхэр гадаргуудаа β-ялтас, гадна гүдгэр гадаргуудаа өөр өөр төрлийн хоёрдогч бүтэц агуулсан байдаг. Бүх ЛБД уургийн β-ялтасын 4-р байрлалд гидрофоб амин хүчил байрлана. Энэхүү програм нь хадгалагдсан дарааллыг ашиглан регуляр илэрхийллийн тусламжтайгаар давталтуудыг таньж, давталт бүрийн 4-р байрлалын гидрофоб амин хүчлийн α-карбоны координатыг автоматаар ялгана. Уг програм нь өгөгдөл дэх нийт давталтуудын 82%-хувийг таньсан бөгөөд эдгээрийг нэмэлт геометрийн шалгууруудаар шалгахад давталтуудын 62% нь программаар зөв танигдсан болох нь батлагдлаа. Энэ програмыг сайжруулж, үр дүнг цаашдын судалгаанд хэрэглэх боломжтой юм.

Түлхүүр үг: **Лейцин-баялаг давталтат уураг, гидрофоб амин хүчил, хадгалагдсан дараалал, MATLAB, регуляр илэрхийлэл**

I. ОРШИЛ

Вирус, бактер, архей болон эукариотоос ялгасан 20000 гаруй уураг Лейцин-баялаг давталт (ЛБД) агуулна. ЛБД агуулсан уураг нь сигнал дамжуулалт, эсийн наалдалт, ДНХ-ийн засварлалт, рекомбинаци, транскрипци, РНХ процессинг, өвчин эсэргүүцэл, эсийн програмчлагдсан үхэл, ургамлын дархлааны хариу үйлдэл, сүүгээр бойжигчдын төрөлхийн дархлааны хариу үйлдэл болон бусад олон уураг-уургийн харилцан үйлчлэлд оролцдог [1-5]. ЛБД агуулсан уургууд болон хажуугийн домейнуудад мутаци болсноос үүдэн альцхаймер, шизофрени, түр зуурын ухаан алдалт, шөнийн сохор, өндгөн эсийн дутуу боловсролт, Паркинсоны өвчин үүсдэг байж болох нь тогтоогдоод байна [6-8].

Лейцин-баялаг давталтын тоо нэг уурагт хоёроос дөчин тавын хооронд байх ба давталт бүр нь ЛБД-т уургийн ангиас хамааран 20-29

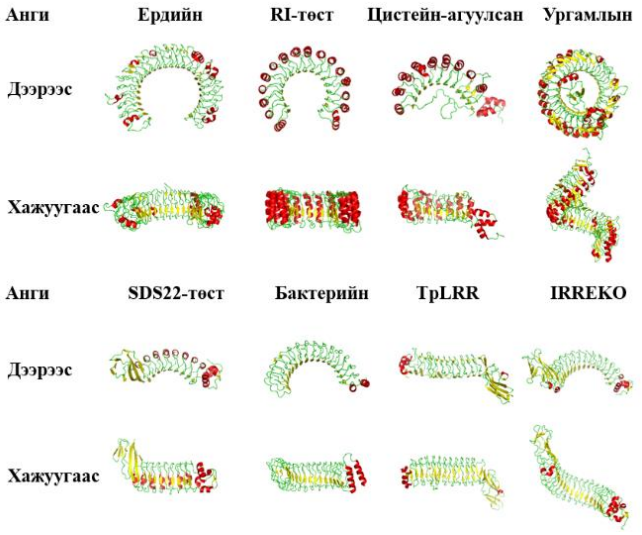


Зураг 1. ЛБД агуулсан Рибонуклеаза ингибитор уургийн (2BNH) дарааллыг давталтуудаар харуулсан байдал. Давталт тус бүрийн 4-р байрлал дээр гидрофоб амин хүчил (L, V, I) байна. Хадгалагдсан дарааллын L-ээр L, V, I амин хүчлүүдийг, X-ээр дурын амин хүчлүүдийг тэмдэглэв.

амин хүчлийн урттай байдаг (Зураг 1) [9, 10]. ЛБД уургийн дарааллыг давталтын тухайн байрлал бүрд олонтоо тааралдах амин хүчлийн дарааллаар нь буюу хадгалагдах дарааллаар нь дүрсэлж болох ба энэ нь ангиас хамаарч өөрчлөгдөхгүй сайн хадгалагдах хэсэг болон хувьсах хэсгээс тогтдог. ЛБД нь хадгалагдсан дарааллын хэлбэр, давталтын урт, хувьсах хэсгийн хэлбэрээрээ найман ангид хуваагддаг (Хүснэгт 1). ЛБД агуулсан уураг нь морины тах, эллипс эсвэл суперхеликс хэлбэрийн бүтэцтэй байх ба дотоод хүнхэр гадаргуудаа β-ялтас, гадаад гүдгэр гадаргуудаа α, 3_{10} болон бусад төрлийн хеликсүүдийг агуулсан байдаг (Зураг 2) [11].

Хүснэгт 1. ЛБД уургийн ангилал. Түүний давталтын урт болон хадгалагдсан дараалал. Хадгалагдсан дараалал дахь тодруулсан хэсэг нь сайн хадгалагдсан хэсэг бол үлдсэн хэсэг нь хувьсах хэсэг юм.

Анги	**Давтал-тын урт**	**Хадгалагдсан дараалал**
Ердийн	20-27	**L**xx**L**x**L**xx**N**x**L**xxLpxxoFxxLxx
RI-төст	28-29	**L**xx**L**x**L**xx**N**x(**L**/**C**)xxxgoxxLxxoLxxxxx
Цистейн-агуулсан	25-27	**L**xx**L**x**L**xx**C**xx**I**TDxxoxxL(a/g)xx(C/L)xx
Ургамлын	23-25	**L**xx**L**x**L**xx**N**x**L**(t/s)GxIPxxLGxLxx
SDS22-төст	21-23	**L**xx**L**x**L**xx**N**(r/k)**I**(r/k)xIE(N/G)LexLxx
Бактерийн	20-22	**L**xx**L**x**V**xx**N**x**L**xxLP(D/E)LPxx
Трепонемо ЛБД	23-25	**L**xx**L**x**L**xxx**L**xxlgxxAFxx(C/N)xx
IRREKO	21	**L**xx**L**x(**L**/**C**)xx**N**x**L**xxLDLxx()xx



Зураг 2. ЛБД уургийн найман ангийн төлөөлөгч болох уургуудын хоёрдогч бүтцийг туузан диаграм ашиглан дээрээс болон хажуугаас PyMOL [11] програмаар дүрслэн үзүүлэв.

II. ӨГӨГДӨЛ

Уургийн бүтцийн өгөгдлийн сангаас (PDB, Protein Data Bank) [12] түлхүүр үгээр эсвэл ЛБД гэдэг нь мэдэгдэж байгаа уургийн дарааллыг ашиглан, ЛБД агуулсан нийт 88 уургийн 189 кристал бүтцийг судалгаанд хэрэглэв (Хүснэгт 2). Эдгээр 189 кристал бүтцийн дан гинжнээс ялгасан лейцин-баялаг давталтуудыг гар аргаар ялгахад нийт 2447 болсон.

Хүснэгт 2. ЛБД уургийн анги, ангид хамаарах уураг, уураг дах давталтын тоо.

Анги	Уургийн тоо	PDB-ийн тоо	Давталтын тоо
Ердийн	42	100	1376
RI-төст	13	19	209
Цистейн-агуулсан	4	16	205
Ургамлын	6	10	227
SDS22-төст	12	31	261
Бактерийн	4	6	73
Трепонемо ЛБД	6	6	81
IRREKO	1	1	15
НИЙТ	**88**	**189**	**2447**

III. АРГА ЗҮЙ

Matlab функцууд

Matlab програмд Биоинформатикийн судалгаанд зориулагдсан Toolbox буюу багц функцууд байдаг бөгөөд биологийн их хэмжээний өгөгдөл дээр гараар хийх үйлдлийг хялбарчлах зорилгоор өргөн ашиглагддаг [13].

`pdbread` функцыг ашиглан уургийн бүтцийг татаж авч, нэг кристал бүтцээс давталтыг хайхдаа регуляр илэрхийллийг ашиглан (`regex` функц) уургийн гинж тус бүрд хайлт хийв.

Регуляр илэрхийлэл

Хайлт хийхийн тулд Matlab програмийн `regexpi(string, expressn)` функцийг ашигласан. Энэ функц үсгийн том жижгийг ялгалгүй хайлт хийх бөгөөд энд байгаа `string` нь дурын урттай, дурын үсгүүдийг агуулж буй задлан шинжлэл хийгдэх дараалал байна. `expressn` нь `string` дотроос хайлт хийж олох тэмдэгтийн онцлог хэв шинжийг заах текст, оператор байна.

ЛБД уурагт хайлт хийхдээ анги тус бүрд харгалзах хадгалагдсан дарааллыг `expressn` болгож, 189 уураг тус бүрийн амин хүчлийн дарааллыг `string` болгож ашигласан.

Геометрийн шалгуур

Регуляр илэрхийллээр хайж олсон үр дүнгүүд дотор бидний олох ёстой давталт биш буюу *false positive* байх тохиолдол байна. Иймээс гарсан үр дүнгүүд үнэн эсэхийг давхар шалгах нэмэлт шалгуур болгон ЛБД уургийн бүтцийн зарчимыг тусгасан зарим геометрийн тооцоолол хийх шаардалагатай болсон.

ЛБД агуулсан уургийн давталтын 4-р байрлал дээрх амин хүчлийн α-карбонаас дараагийн давталтын 4-р байрлал дээрх амин хүчлийн α-карбон хүртэлх зай нь 4.5-аас 5.5-ийн хооронд байдаг. Үүнийг зайн буюу геометрын нэгдүгээр шалгуур болгож ашиглав. A(x_1, y_1, z_1), B(x_2, y_2, z_2) цэгийн хоорондох зай нь

$$|AB| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (1)$$

гэж бодогдоно.

Давталт бүрийн 4-р байрлал дээрх амин хүчлүүд нь зэрэгцээ байрлаж уургийн гидрофоб корыг үүсгэх учраас эдгээр амин хүчлийн β-карбон нь уургийнхаа дотор тал руу буюу дотоод хүнхэр гадаргуу талаасаа гадаад гүдгэр гадаргуу тал руугаа чиглэсэн байдаг [5]. Иймд эдгээр амин хүчлүүдийн α-карбонаас β-карбоныг холбосон векторын чиглэлийг геометрын хоёрдугаар шалгуур болгон ашиглав. Эхний давталтын 4-р байрлал дээрх амин хүчлийн α- ба β-карбоныг холбосон вектор уургийнхаа дотоод тал руу чиглэсэн байна гэж үзвэл дараагийн давталтуудын ийм векторууд уургийнхаа дотоо тал руу чиглэсэн байхын тулд эхний вектортай 90°-аас өнцөг үүсгэж байх ёстой болно. $\overrightarrow{AB} = (x_i, y_i, z_i)$ болон $\overrightarrow{CD} = (x_j, y_j, z_j)$ векторуудын хоорондох өнцөг нь атомуудын координатаар

$$\alpha = arccos\left(\frac{x_j x_k + y_j y_k + z_j z_k}{\sqrt{x_j^2 + y_j^2 + z_j^2}\sqrt{x_k^2 + y_k^2 + z_k^2}}\right) \quad (2)$$

гэж бодогдоно.

IV. ҮР ДҮН БА ХЭЛЭЛЦҮҮЛЭГ

ЛБД уургийн давталтыг регуляр илэрхийллээр таних энэхүү програмыг нийт 88 уургийн 189 кристал бүтцийн 2447 давталтын танилт хийлгэж шалгалаа. Үр дүнд нийт давталтуудын 82.67% буюу 2023 давталтыг таньсан (Хүснэгт 3) ба энэ үр дүнгийн зөв эсэхийг шалгахын тулд зай, өнцгийн шалгуурыг нэмж оруулахад нийт давталтуудын 62.32% буюу 1525 давталт зөв танигдсан байна (Хүснэгт 4).

Хүснэгт 3. Бодит давталтын тоо, регуляр илэрхийллээр танигдсан давталтын тоо, хувь

Анги	Давталтын тоо	Үр дүн	Хувь (%)
Ердийн	1376	1152	83.72
RI-төст	209	159	76.08
Цистейн-агуулсан	205	152	74.15
Ургамлын	227	202	88.99
SDS22-төст	261	205	78.54
Бактерийн	73	70	95.89
Трепонемо ЛБД	81	69	85.19
IRREKO	15	14	93.33
НИЙТ	**2447**	**2023**	**82.67**

Хүснэгт 4. Бодит давталтын тоо, зай болон өнцгийн шалгуурыг хангасан давталтын тоо, хувь

Анги	Давталтын тоо	Үр дүн	Хувь(%)
Ердийн	1376	837	60.83
RI-төст	209	130	62.20
Цистейн-агуулсан	205	77	37.56
Ургамлын	227	190	83.70
SDS22-төст	261	174	66.67
Бактерийн	73	70	95.89
Трепонемо ЛБД	81	33	40.74
IRREKO	15	14	93.33
НИЙТ	**2447**	**1525**	**62.32**

Зөвхөн регуляр илэрхийллээс гарсан үр дүнгээс үзэхэд ЛБД уургийн ангиас хамааран харилцан адилгүй давталтыг илрүүлж байна. Хамгийн сайн танигдсан анги болох бактерийн анги нь 6 кристал бүтцийн нийт 73 давталттай байснаас 70 давталтыг таньсан. Эдгээр давталтуудын зай болон өнцөг нь шалгуурыг хангаж байгаа нь давталтыг яг зөв таньсан байгааг илэрхийлнэ. Бактерийн ангийн давталтын хадгалагдсан дараалал нь бусад ангиудаас илүү тогтвортой байдаг ба давталтын урт нь 20-22 байдаг учраас шалгуур хангасан үр дүн сайн гарчээ. Дараагийн сайн үр дүнтэй гарсан анги нь IRREKO юм. Энэ ангид бид 15 давталттай ганц кристал бүтцийг шалгаж үзсэн. Ганцхан кристал бүтцэд шалгалт хийгдсэн учраас алдаагүй үр дүн гарсан байна.

Харин цистейн-агуулсан болон трепонемо ЛБД ангийн хувьд регуляр илэрхийллээс гарсан үр дүнгийн талаас их хувь нь шалгуур хангаагүй давталтууд буюу буруу үр дүн гаргасан байна. Үүний учир нь цистейн-агуулсан ангийн хадгалагдсан дараллын 9-р байрлал дээр цистейн агуулдаг гэсэн байгаа боловч үүний оронд бүх гидрофоб амин хүчлүүд мөн өвөрмөц амин хүчлүүд болох глицин, пролиныг агуулахын зэрэгцээ цэнэггүй туйлт амин хүчлүүдийг агуулдаг болох нь гар аргаар давталтыг илрүүлэх үед ажиглагдсан. Тиймээс энэ шалгуурыг бичихэд дурын амин хүчил гэж тодорхойлж өгсөнөөс ялгаагүй үр дүн гаргаж байгаа нь давталтыг буруу таньж байгаа нэгэн шалтгаан болно. ТрепонемоЛБД ангийн хувьд бүтэц болон дараалалд агуулагдаж байгаа гидрофоб амин хүчлийн тооны хувьд бусад ангиас ялгаатай байдаг. Энэ ангийн уургуудын бүтэц нь гадаад гүдгэр гадаргуудаа мөн дотоод хүнхэр гадаргуу аль алиндаа β-ялтасыг агуулсан байдаг. Ийм олон гидрофоб амин хүчлүүд байгаа нь програмыг хэтэрхий олон давталт илрүүлэхэд хүргэж байх магадлалтай байна. ЛБД уургийн бусад ангиудын хувьд уурагт агуулагдаж байгаа давталтуудын урт хэлбэлзэл ихтэй байдаг учраас түүнийг регуляар илэрхийллээр илрүүлж, тодорхойлоход хэцүү байгаа байдал ажиглагдаж байна.

V. ДҮГНЭЛТ

ЛБД уургийн анги бүрийн хадгалагдсан дарааллыг ашиглан регуляр илэрхийлэл зохиож 88 уургийн 189 кристал бүтцийн 2447 ЛБД-ыг таниулах оролдлого хийлээ. 2447 давталтаас 2023-ийг буюу нийт 82.67 хувийн давталт гэж илрүүлэв. ЛБД уургийн бүтцийн геометрын шинж чанаруудыг нэмэлт шалгуур болгон шалгах 62.32 хувь нь зөв танигдлаа. Үүнээс үзвэл 17.33 хувийг огт танихгүй 20.35 хувийг буруу таньж байна. Энэ програмаар ялган авсан атомын координатуудыг ЛБД-ийн геометрын параметрүүд (ерөнхий радиус, суперхеликсийн алхам зэрэг)-ийн тооцоонд өгөгдөл болгон хэрэглэж, тооцооны ажлын автоматжуулах боломжтой юм.

НОМ ЗҮЙ

[1] Matsushima, N., Miyashita, H., (2010). “A nested Leucine Rich Repeat Domain: The Precursor of LRRs is a ten or eleven residue motif.” Microbiology **10:**235-245.

[2] Kobe B., and Deisenhofer J., (1994). “Leucine Rich Repeats: a versatile binding motif.” Trends in Biochemical Science 9: 415-421.

[3] Kobe B., and Deisenhofer J., (1995). “Proteins with leucine-rich repeats.” Curr. Opin. Struct. Biol. 5(3):409-416.

[4] Kobe B., and Kajava A.V. (2000). “When protein foldig is simplified to protein coiling: the continuum of solenoid protein” TIBS 25: 509-515.

[5] Enkhbayar, P., Kamiya, M., et al. (2004). “Structural principles of leucine-rich repeat (LRR) proteins.” Proteins 54(3): 394-403.

[6] Kajava, A.V., (1998). “Structural diversity of leucine-rich repeat proteins.” J.Mol.Biol 277(3): 519-527.

[7] Bella J., Hindle. K.L., et al. (2008). “The leucine-rich repeat structure.” Cell Mol. Life Sci. 65(15): 2307-2333.

[8] Matsushima, N., Tachi, N., Enkhbayar, P., (2005). “Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases.” Cell Mol. Life Sci. 62(23): 2771-2791.

[9] Miyashita, H., Kuroki, Y., et al. (2013). “Horizontal gen transfer of plant-specific leucine-rich repeats between plants and bacteria.” Natural Sci. 5(5):

[10] Kobe, B., Kajava, A.V., (2001). “The leucine rich repeats as a protein recognition motif.” Curr.Opin.Struct.Biol. 11: 725-732.

[11] Delano, W,.L., “The PyMOL Molecular Graphics System” 2002

[12] Berman. H.M., Westbrook. J., et al. (2000). “The protein data bank” Oxford Journals. 28(1): 235-242.

[13] URL: http://www.mathworks.com/