11-ДӘРІС

Биологиялық деректерді талдау

Пән: CHE6281 – Инженерлік биотехнология | Дәріс түрі: теориялық | Ұзақтығы: 2 академиялық сағат

Дәрістің мақсаты

Биотехнологиялық зерттеу және өндіріс барысында алынатын биологиялық деректерді жинау, ұйымдастыру, сапасын бақылау, статистикалық талдау, визуализациялау және ғылыми тұрғыдан интерпретациялау қағидаларын меңгерту.

Күтілетін оқу нәтижелері

- Биологиялық деректердің негізгі түрлерін жіктейді.
- Өлшеу қатесі, биологиялық өзгергіштік және деректер сапасының айырмашылығын түсіндіреді.
- Сипаттамалық статистика мен гипотезаларды тексерудің негізгі әдістерін таңдайды.
- Корреляция мен регрессияны ажыратады.
- Көптік салыстырулар мен репликацияның маңызын түсіндіреді.
- Биотехнологиялық көрсеткіштерді есептеп, нәтижелерді график және кесте түрінде ұсынады.

Дәріс жоспары

- Биологиялық деректер және олардың көздері
- Деректерді ұйымдастыру және сапасын тексеру
- Сипаттамалық статистика
- Гипотезаларды тексеру және статистикалық қателер
- Корреляциялық және регрессиялық талдау
- Эксперимент дизайны және көптік салыстырулар
- Омикалық және биоинформатикалық деректер
- Визуализация, репродукция және деректер этикасы
- Инженерлік есептер мен өндірістік кейс

Кіріспе

Инженерлік биотехнологияда деректер биореактор датчиктерінен, зертханалық талдаулардан, жасуша өсіру тәжірибелерінен, фермент белсенділігін өлшеуден, ген экспрессиясын зерттеуден және өнім сапасын бағалаудан алынады. Деректерді дұрыс талдау технологиялық режимдерді салыстыруға, өндірістік ауытқуларды анықтауға және ғылыми тұжырымдарды негіздеуге мүмкіндік береді. Статистикалық маңыздылық биологиялық немесе өндірістік маңыздылықпен әрдайым бірдей емес.

1. Биологиялық деректердің түрлері

1.1. Дерек көздері

Дерек түрі	Мысал	Өлшем бірлігі немесе пішімі
Үздіксіз сандық	Биомасса, температура, pH	г/л, °C, шартты бірлік
Дискретті сандық	Колония саны, жасуша саны	дана
Категориялық	Өндірістік партия, генотип	санат
Уақыттық қатар	DO немесе pH өзгерісі	мән + уақыт белгісі
Омикалық	РНҚ экспрессиясы, протеомика	санақ/қарқындылық матрицасы

1.2. Тәуелсіз және тәуелді айнымалылар

Тәуелсіз айнымалы зерттеуші өзгерткен немесе салыстырған фактор болуы мүмкін (мысалы, температура); тәуелді айнымалы – оның әсерін бағалау үшін өлшенетін нәтиже (мысалы, өнім концентрациясы). Бақылаусыз факторлар нәтижеге ықпал етуі мүмкін. Оларды тіркеу және талдауда ескеру маңызды.

1.3. Биологиялық және техникалық қайталанулар

Биологиялық қайталану – бір-бірінен тәуелсіз биологиялық үлгі немесе эксперименттік бірлік. Техникалық қайталану – сол үлгінің өлшеуін қайталау. Үш техникалық өлшеуді үш тәуелсіз биологиялық тәжірибе деп есептеуге болмайды: мұндай қате псевдорепликацияға әкеледі.

2. Деректерді ұйымдастыру және сапаны бақылау

2.1. Деректер кестесінің құрылымы

Көптеген талдауларда әр жол – бір бақылау, әр баған – бір айнымалы. Айнымалылардың атауы, өлшем бірлігі, үлгі идентификаторы, эксперимент күні, партия және қолданылған әдіс бөлек тіркелгені дұрыс. Деректер сөздігі әр бағанның мағынасын түсіндіреді.

2.2. Жетпейтін мәндер және шеткі бақылаулар

Жетпейтін мәндердің себебі техникалық ақау, үлгінің жарамсыздығы немесе тіркеу қатесі болуы мүмкін. Оларды автоматты түрде нөлге алмастыруға болмайды. Шеткі бақылау өлшеу қатесі де, нақты биологиялық ерекшелік те болуы мүмкін; сондықтан алып тастау критерийі алдын ала анықталып, құжатталуы тиіс.

2.3. Нормализация және стандарттау

Әртүрлі масштабтағы өлшемдерді салыстыру үшін нормализация немесе стандарттау қолданылады. Z-балл: $z = (x - \bar{x})/s$, мұнда $\bar{x}$ – орташа мән, s – стандартты ауытқу. Омикалық деректердегі нормализация әдісі платформа мен эксперимент дизайнына байланысты таңдалады; барлық дерекке бір ғана тәсіл сәйкес келмейді.

3. Сипаттамалық статистика

3.1. Орталық тенденция көрсеткіштері

Арифметикалық орта: $\bar{x} = \Sigma xi/n$. Медиана – реттелген қатардың орталық мәні. Орта симметриялы деректерді сипаттауға ыңғайлы; ал медиана шеткі мәндердің ықпалы жоғары болған жағдайда орнықтырақ болуы мүмкін.

3.2. Шашырау көрсеткіштері

Таңдамалық дисперсия: $s^2 = \Sigma(xi - \bar{x})^2/(n - 1)$; стандартты ауытқу $s = \sqrt{s^2}$. Квартильаралық ауқым $IQR = Q3 - Q1$. Стандартты қате $SEM = s/\sqrt{n}$, бірақ SEM жеке бақылаулардың шашырауын көрсетпейді. SD мен SEM-ді графикте нақты ажыратып жазу керек.

3.3. Сенімділік аралығы

Сенімділік аралығы параметрді бағалаудағы белгісіздікті көрсетеді. Орта үшін жуық 95% сенімділік аралығы $\bar{x} \pm t^* \times s/\sqrt{n}$ түрінде есептеледі; t^* мәні

еркіндік дәрежесіне тәуелді. Бұл интервалды жеке болашақ өлшеудің ықтимал аралығымен шатастыруға болмайды.

Көрсеткіш	Не сипаттайды?	Ескерту
Орташа мән	Орталық деңгей	Шеткі мәндерге сезімтал
Медиана	Қатардың ортасы	Қиғаш таралымда пайдалы
SD	Бақылаулардың өзгергіштігі	Бірлігі бастапқы дерекпен бірдей
SEM	Орта бағасының дәлдігі	SD орнына қолданылмайды
95% CI	Бағалаудың белгісіздігі	Үлгі көлемі мен модельге тәуелді

4. Статистикалық гипотезаларды тексеру

4.1. Нөлдік және баламалы гипотезалар

H0 әдетте айырмашылық немесе байланыс жоқ деген тұжырымды, H1 – баламалы тұжырымды білдіреді. p-мәні H0 дұрыс деп алынған жағдайда байқалғандай немесе одан да шеткі нәтиженің ықтималдығын сипаттайды. Ол H0 гипотезасының дұрыс болу ықтималдығы емес.

4.2. I және II текті қателер

I текті қате – дұрыс H0 гипотезасын теріске шығару; оның жоспарланған ықтималдығы α арқылы бақыланады. II текті қате – жалған H0 гипотезасын теріске шығара алмау; статистикалық қуат $1 - \beta$ арқылы сипатталады. Қуат әсер шамасына, үлгі санына, өзгергіштікке және тестке байланысты.

4.3. Тест таңдау

Зерттеу жағдайы	Мүмкін әдіс	Негізгі шарт
Екі тәуелсіз топтың орташа мәндері	Welch t-тесті	Тәуелсіз бақылаулар; үлестірім мен көлемді бағалау
Жұпталған өлшемдер	Жұпталған t-тест	Бақылаулардың жұпталуы
Үш немесе одан көп топ	ANOVA немесе Welch ANOVA	Дизайн мен дисперсия ерекшеліктерін ескеру

Екі категориялық айнымалы	χ^2 немесе Fisher дәл тесті	Жиіліктер және күтілетін сан
Реттік/айқын қиғаш деректер	Рангтік әдістер	Тәуелсіздік және тесттің нақты болжамдары

Тест таңдау тек «нормалдық тестінің p-мәніне» тәуелді болмауы керек. Эксперимент дизайны, бақылаулардың тәуелсіздігі, үлгі көлемі және бағаланатын әсер маңызды.

5. Корреляция және регрессиялық талдау

5.1. Корреляция

Пирсон корреляция коэффициенті r екі сандық айнымалының сызықтық байланысының бағыты мен күшін сипаттайды; r −1 мен +1 аралығында болады. Спирменнің рангтік корреляциясы монотонды байланыстарды бағалауда пайдалы. Корреляция себеп-салдарлық байланысты дәлелдемейді.

5.2. Сызықтық регрессия

Қарапайым сызықтық модель: $Y = \beta0 + \beta1X + \varepsilon$. Мұнда β0 – бос мүше, β1 – X бір бірлікке өзгергендегі Y-тің күтілетін өзгерісі, ε – модель ескермеген ауытқу. Регрессия нәтижесінде коэффициент, оның сенімділік аралығы, p-мәні және модельдің жарамдылық көрсеткіштері бірге берілуі тиіс.

5.3. Көптік регрессия және шатастырушы факторлар

Көптік регрессияда бірнеше түсіндірмелі айнымалы қарастырылады: $Y = \beta0 + \beta1X1 + \beta2X2 + \ldots + \varepsilon$. Мысалы, ферментация өнімділігін температура, pH және субстрат концентрациясымен бірге модельдеуге болады. Факторлардың өзара байланысы (мультиколлинеарлық), қалдықтардың құрылымы және модельден тыс экстраполяция ескеріледі.

6. Эксперимент дизайны және көптік салыстырулар

6.1. DOE қағидалары

Тәжірибелерді жоспарлау (DOE) факторлардың жеке әсерлерін және өзара әрекеттесуін бағалауға мүмкіндік береді. Рандомизация жүйелі ығысуды азайтады, блоктау белгілі өзгергіштік көздерін бөледі, ал тәуелсіз қайталанулар кездейсоқ ауытқуды бағалауға көмектеседі.

6.2. Көптік тексеру мәселесі

Бір уақытта көптеген гипотеза тексерілсе, кездейсоқ «маңызды» нәтижелер саны өсуі мүмкін. Bonferroni тәсілі отбасылық қате ықтималдығын шектеуге бағытталған; Benjamini–Hochberg процедурасы жалған ашылымдар үлесін (FDR) бақылауға арналған. Әдісті таңдау зерттеу мақсатына және тесттер жиынтығына байланысты.

6.3. Әсер шамасы

Тек p-мәнін көрсету жеткіліксіз. Топтар арасындағы орташа айырмашылық, қатынас көрсеткіштері немесе стандартталған әсер шамасы нәтижені практикалық тұрғыдан түсіндіруге көмектеседі. Әсер шамасын сенімділік аралығымен бірге ұсыну ұсынылады.

7. Омикалық және биоинформатикалық деректер

7.1. Геномика және транскриптомика

Геномдық талдауда нуклеотидтік тізбектер, варианттар және генотиптер зерттеледі. Транскриптомикада гендердің экспрессия деңгейлері салыстырылады. Деректерді өңдеудің жалпы кезеңдері: сапаны бағалау, тиісті сәйкестендіру/санақ, нормализация, статистикалық модельдеу, көптік салыстыруды түзету және биологиялық интерпретация.

7.2. RT-qPCR деректерін талдау

RT-qPCR кезінде Cq мәні бастапқы нысана мөлшерімен кері байланысты. Салыстырмалы экспрессияны есептеуде $\Delta Cq = Cq(\text{нысана}) - Cq(\text{референс})$; $\Delta\Delta Cq = \Delta Cq(\text{тәжірибе}) - \Delta Cq(\text{бақылау})$; салыстырмалы өзгеріс $\approx 2^{(-\Delta\Delta Cq)}$. Бұл тәсіл нысана мен референс амплификация тиімділіктері жеткілікті ұқсас болғанда және референс ген тұрақты болғанда орынды. Биологиялық репликаттар арасындағы талдау дұрыс эксперименттік бірліктерге негізделуі тиіс.

7.3. Протеомика және метаболомика

Протеомика ақуыздардың мөлшері мен қасиеттерін, метаболомика жасушалық метаболиттер жиынтығын зерттейді. Бұл деректерде өлшеу диапазоны, партиялық әсер, жетпейтін мәндер және көпөлшемділік жиі кездеседі. Негізгі компоненттер талдауы (PCA) үлгілер арасындағы жалпы өзгергіштікті визуализациялауға көмектеседі, бірақ өздігінен себепті немесе статистикалық маңыздылықты дәлелдемейді.

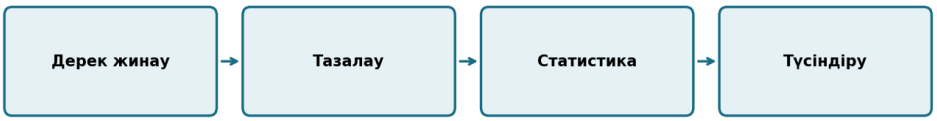
7.4. Деректерді біріктіру

Көпомикалық зерттеулер ген экспрессиясы, ақуыз мөлшері және метаболиттерді бірге қарастыруға мүмкіндік береді. Әртүрлі платформалардың масштабтары мен қателері бөлек болғандықтан, деректерді сәйкестендіру, нормализация және тәуелсіз тексеру қажет.

8. Деректерді визуализациялау және ғылыми есеп беру

- Нүктелік диаграмма екі сандық айнымалының байланысын көрсетеді.
- Қорапша диаграмма медиана, квартильдер және таралу ерекшеліктерін көрсетеді.
- Уақыттық график процестің динамикасын сипаттайды.
- Жылу картасы (heatmap) көпөлшемді деректердегі үлгілерді көрсетуге көмектеседі.
- График осьтерінде өлшем бірлігі, үлгі саны және белгісіздік көрсеткіштері берілуі керек.

Графикте у-осьті негізсіз қиып көрсету немесе SD мен SEM-ді шатастыру нәтиженің көрнекі түсіндірмесін бұрмалайды. Зерттеу есебінде әдістер, критерийлер, деректерді өңдеу ережелері, бағдарламалық құралдар және статистикалық болжамдар айқын жазылуы тиіс.



1-сурет. Биологиялық деректерді талдау бойынша оқу сызбасы.

9. Деректерді сақтау, қайта өндіру және этика

Ғылыми қайта өндіру үшін бастапқы деректерді, өңдеу қадамдарын, талдау кодының нұсқасын және метадеректерді сақтау маңызды. Деректерді басқару жоспары файл құрылымын, резервтік көшіруді, қолжетімділікті және сақтау мерзімін анықтайды. FAIR қағидалары деректердің табылуын, қолжетімділігін, үйлесімділігін және қайта қолданылуын жақсартуға бағытталған.

Адамдарға қатысты генетикалық және медициналық деректерді бөлісу кезінде ақпараттандырылған келісім, құпиялылық және қолданылатын құқықтық

талаптар ескеріледі. Ашық ғылым қағидалары құпия деректерді шектеусіз жариялау міндетін білдірмейді.

10. Инженерлік есептер

10.1. Орташа мән және стандартты ауытқу

Мысал: үш тәуелсіз биореакторда өнім концентрациясы 8, 10 және 12 г/л болды. Орташа мән $\bar{x} = (8 + 10 + 12)/3 = 10$ г/л. Таңдамалық дисперсия $s^2 = [(8-10)^2 + (10-10)^2 + (12-10)^2]/(3-1) = 4$ (г/л)²; SD = 2 г/л; SEM = $2/\sqrt{3} \approx 1{,}15$ г/л.

10.2. Өнім шығымын есептеу

Мысал: 75 г субстрат жұмсалып, 30 г мақсатты өнім түзілді. Өнім шығымы $Y_{P/S} = 30/75 = 0{,}40$ г/г. Егер тәуелсіз партиялар болса, әр партияның шығымын жеке есептеп, содан кейін олардың таралуын бағалау қажет.

10.3. RT-qPCR мысалы

Бақылау тобында $\Delta Cq = 5$, тәжірибе тобында $\Delta Cq = 3$ болсын. $\Delta\Delta Cq = 3 - 5 = -2$. Салыстырмалы экспрессия = $2^{(-(-2))} = 4$ есе. Бұл тек шартты оқу есебі: нақты зерттеуде тиімділік, референс ген тұрақтылығы және тәуелсіз биологиялық қайталанулар тексеріледі.

10.4. Өндірістік уақыттық қатар

Биореактордағы еріген оттегі көрсеткіші бірнеше сағат ішінде төмендеп, кейін қайта көтерілді делік. Бұл өзгерісті бірден «жасушалардың өсуі тоқтады» деп түсіндіруге болмайды. Аэрация, араластыру, сенсор калибрленуі, биомасса және субстрат деректерін бірге салыстыру қажет.

11. Өндірістік кейс: екі биопроцесті салыстыру

Кәсіпорын А және В қоректік орталарын салыстырады. Әр орта үшін бірнеше тәуелсіз өндірістік партияда өнім концентрациясы, ферментация уақыты және субстрат шығыны тіркеледі. А ортасында өнім концентрациясы жоғары болуы мүмкін, бірақ процесс ұзаққа созылса, көлемдік өнімділік төмендеуі ықтимал. Сондықтан шешім тек бір көрсеткішке емес, өнімділікке, шығымға, сапаға, белгісіздікке және шығынға сүйенуі керек.

- Эксперименттік бірлік ретінде нені алу керек?
- Қандай сипаттамалық статистиканы көрсету орынды?

- Екі топты салыстырғанда қандай статистикалық болжамдарды тексеру қажет?
- Өндірістік шешім үшін p-мәнінен басқа қандай көрсеткіштер керек?

Қорытынды

Биологиялық деректерді талдау – биотехнологиялық зерттеу мен өндірістік шешімдердің ғылыми негізі. Сапалы талдау деректерді дұрыс жинаудан, тәуелсіз қайталануларды анықтаудан және өлшеу қателерін бақылаудан басталады. Статистикалық тест, корреляция, регрессия және көпөлшемді талдау әдістері зерттеу мақсатына сәйкес таңдалуы тиіс. Нәтижелерді түсіндіруде әсер шамасы, белгісіздік, биологиялық мән және қайта өндіру мүмкіндігі бірге қарастырылады.

Бақылау сұрақтары

1. Биологиялық деректердің негізгі түрлерін атаңыз.
2. Биологиялық және техникалық қайталанулар қалай ажыратылады?
3. Псевдорепликация дегеніміз не?
4. Жетпейтін мәндер мен шеткі бақылауларды қалай бағалау керек?
5. Орташа мән мен медиананың айырмашылығы қандай?
6. SD, SEM және сенімділік аралығы нені сипаттайды?
7. p-мәнін қалай дұрыс түсіндіруге болады?
8. I және II текті қателер дегеніміз не?
9. Welch t-тесті қандай жағдайда қолданылады?
10. Корреляция себеп-салдарды неге дәлелдемейді?
11. Көптік регрессияда қандай факторлар ескеріледі?
12. FDR түзетуі не үшін қажет?
13. RT-qPCR үшін ΔΔCq тәсілінің негізгі шарттары қандай?
14. PCA қандай мақсатта қолданылады?
15. FAIR қағидалары мен ғылыми деректердің қайта өндірілуі неліктен маңызды?