CALIBRACIÓN MEJORADA DE UN MODELO DE FERMENTACIÓN EN SUSTRATO SÓLIDO

Pedro Saa, Martín Cárcamo, Javiera López, J. Ricardo Pérez-Correa, Claudio Gelmi

INTRODUCCIÓN

Los procesos en biotecnología inicialmente se estudian por lo general a escala de laboratorio en sistemas batch, en que el análisis resulta complicado debido al carácter dinámico de estos. Los modelos que describen este tipo de sistemas, son de tipo dinámico y normalmente contienen un gran número de parámetros que es necesario calibrar a partir de un conjunto limitado de experimentos.

El proceso general para identificar un modelo consiste en encontrar el set de parámetros óptimos que permite reproducir la dinámica observada del sistema de la forma más precisa posible. Este proceso, conocido como regresión o calibración - básicamente una optimización - puede validarse contrastando los resultados del modelo con nuevos datos experimentales. En modelos biotecnológicos complejos es conveniente utilizar métodos de optimización global (GO) para encontrar el set de parámetros óptimos. Estos métodos se traducen en códigos robustos que pueden localizar el óptimo global en un número razonable de iteraciones y son capaces de manejar el ruido y/o las discontinuidades de la función objetivo. Los algoritmos estocásticos – un tipo de GO – tratan la función objetivo como una caja negra, es decir, como una simple relación entre las entradas y salidas. Dentro de ellos, los algoritmos metaheurísticos son especialmente útiles y cada vez más utilizados en biotecnología. Ellos consideran procesos iterativos que encuentran eficientemente soluciones

Pontificia Universidad Católica de Chile, Departamento de Ingeniería Química y Bioprocesos.
De izquierda a derecha: **Martin Cárcamo, Javiera López y Pedro Saa,** alumnos.
No aparecen en la foto los profesores, **Claudio Gelmi y Ricardo Pérez-Correa.**

Contacto: Ricardo Pérez-Correa - perez@ing.puc.cl
Claudio Gelmi - cgelmi@ing.puc.cl

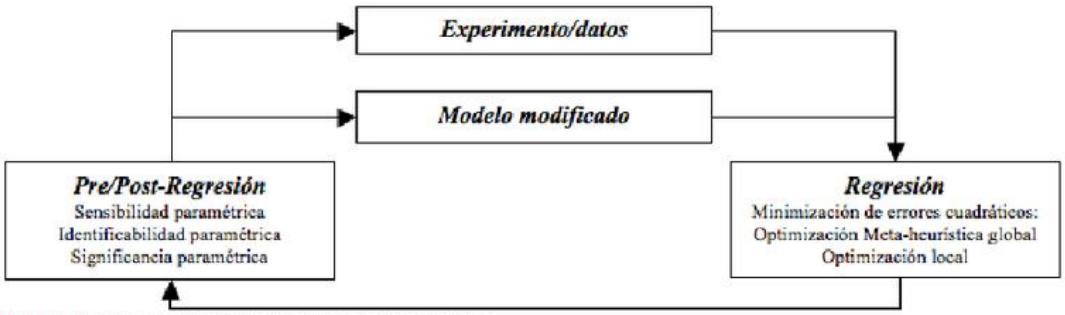


Figura 1: Ciclo iterativo de vinculación entre el modelo y los datos

cercanas al óptimo, combinando de forma adecuada diferentes estrategias de aprendizaje para explorar los espacios de búsqueda. Estos métodos son relativamente fáciles de codificar, lo que los hace apropiados para una amplia variedad de problemas.

Sin embargo, un buen ajuste a los datos experimentales no asegura que se haya encontrado un modelo confiable. Los modelos con un gran número de parámetros tienden a ajustarse mejor a los datos experimentales, pero generan estimaciones poco confiables de los parámetros. Por lo tanto, la calidad de la estimación de parámetros, en términos de precisión, se debe comprobar antes de llegar a una interpretación significativa de los resultados. Los procedimientos necesarios para probar la calidad de la estimación de parámetros se llaman diagnósticos de pre/post-regresión y abarcan varios métodos:

- Sensibilidad paramétrica: como perturbaciones en el valor de un parámetro dado afectan las variables medidas.
- Significancia paramétrica: determinación de la confiabilidad de la estimación del parámetro (intervalo de confianza).
- Identificabilidad paramétrica: detección de la correlación cruzada entre los parámetros.

Por ejemplo, la validación del modelo requiere que los parámetros sean identificables. Luego, si los diagnósticos de post-regresión no son satisfactorios, el modelo tiene que ser modificado y/o los datos experimentales deben repetirse o complementarse con nuevos datos. A su vez, los métodos de diagnóstico pueden dar valiosas recomendaciones para la modificación de montajes experimentales con el fin de aumentar el contenido de información de los datos. Los diagnósticos post-regresión son similares en función a los diagnósticos previos a la regresión, ya que el proceso de vinculación de datos a los modelos es inherentemente iterativo (Fig.1).

En el siguiente trabajo se presenta la calibración de un modelo de crecimiento en sustrato sólido descrito por Gelmi *et al.* [4]. Este tipo de cultivo consiste en el crecimiento de microorganismos, generalmente hongos, en medios sólidos o semi-sólidos en ausencia de agua libre. Las principales aplicaciones de este tipo de cultivos van desde la producción de alimentos hasta la obtención de enzimas de alto valor comercial. El algoritmo de optimización usado está basado en un método de búsqueda dispersa implementado en Matlab® (SSm), el cual ha mostrado resultados prometedores para la solución de problemas de optimización combinatoria y no lineal [2]. Para mejorar la calibración del modelo se utilizaron herramientas de diagnóstico de pre/post-regresión, las cuales consideran análisis de sensibilidad, identificabilidad y significancia de los parámetros del modelo.

EXPERIMENTACIÓN

En este trabajo se utilizó el modelo descrito por Gelmi *et al.* [4] para predecir los cambios de biomasa y producción de giberelina por el hongo filamentoso Gibberella fujikuroi en cultivos de sustrato sólidos. El modelo consiste en ocho ecuaciones diferenciales ordinarias acopladas y catorce parámetros desconocidos. Las variables de estado utilizadas para describir el crecimiento de este microorganismo fueron: biomasa activa, biomasa total, urea, intermediario nitrogenado, almidón, ácido giberélico (GA_3), oxígeno y dióxido de carbono. Se utilizaron cuatro condiciones experimentales. El método de optimización global utilizado en este trabajo fue el algoritmo de búsqueda dispersa. Para validar el modelo se analizó la sensibilidad de los parámetros, la identificabilidad y la significancia estadística de cada uno de ellos.

Análisis de Sensibilidad

El análisis de sensibilidad de los parámetros define en qué medida las variables de estado del modelo son afectadas por las variaciones en los valores de los parámetros. Los parámetros no sensibles son aquellos que no influyen en las variables de estado. En los modelos dinámicos, la sensibilidad varía con el tiempo y por lo tanto los parámetros pueden ser no sensibles en ciertos intervalos de tiempo y sensibles en otros. En general, si un parámetro no es sensible, puede fijarse o bien eliminarse del modelo.

Análisis de Identificabilidad

La identificabilidad calcula la matriz de correlación de los parámetros del modelo para lograr determinar si los parámetros ajustados son localmente identificables para un intervalo de tiempo dado. Si dos parámetros están altamente correlacionados, se dice que son *a priori* identificables y ellos afectan las variables medidas de misma forma. En general, coeficientes de correlación (κ_{ij}) mayor a 0,95 indican una alta correlación. En estos casos los parámetros no pueden ser determinados de manera única.

Análisis de Significancia

La significancia y los intervalos de confianza son calculados después que los parámetros han sido ajustados a los datos experimentales. Para establecer la significancia estadística de las estimaciones normalmente se calcula el *t-value* correspondiente para cada parámetro y se compara con la distribución *t-student* de referencia con (n-p) grados de libertad al 95% de confianza. Valores grandes de *t-values* indican alta confianza y generan intervalos de confianza pequeños.

RESULTADOS Y DISCUSIÓN

En esta sección se muestran los principales resultados de los diferentes métodos de diagnóstico utilizados en este trabajo junto con las modificaciones realizadas sobre la base de dicho análisis. Además, se contrasta el rendimiento de la técnica de optimización global utilizada en este trabajo, con las técnicas estándar utilizadas en publicaciones anteriores como en el caso de Araya *et al.* [1] quienes usaron Colocaciones Ortogonales en Elementos Finitos (OCFE) para la calibración de los parámetros.

Análisis de Identificabilidad

El análisis de identificabilidad mostró que entre los 14 parámetros hay varios pares que están altamente correlacionados. En primer lugar, se encontró que la tasa máxima de crecimiento específico μ_{max} estaba altamente correlacionada con la constante de inhibición k_N en todas las condiciones de cultivo ($\kappa_{ij} > 0{,}98$). El procedimiento habitual para hacer frente a esta situación es mantener constante uno de los parámetros correlacionados. En nuestro caso, fijamos μ_{max} dado que contamos con valores confiables reportados por Gelmi *et al.* [3]. También, se encontró una correlación significativa en los parámetros asociados a los niveles de almidón, CO_2 y O_2 en las diferentes condiciones de cultivo ($\kappa_{ij} > 0{,}95$). Los pares de términos altamente correlacionadas fueron: YX/S y mS, YX/CO_2 y mCO_2, YX/O_2 y mO_2, los cuales corresponden a los rendimientos de biomasa (YX) y las tasas de mantención (m) para cada compuesto. Sin embargo, en estos casos no se cuenta con datos reportados en literatura acerca del valor de estos parámetros. En consecuencia, otros experimentos han de ser planificados en cada estado a fin de identificar cada uno de los parámetros de forma independiente. En este estudio se fijaron arbitrariamente YX/S, YX/CO_2 y YX/O_2 a los valores obtenidos por Gelmi *et al.* [4].

Análisis de Sensibilidad

Nuestro análisis de sensibilidad mostró que la tasa de degradación de GA_3 (k_p) no exhibe gran sensibilidad en las variables de estado para cada condición de cultivo. Por lo tanto, con el fin de facilitar los esfuerzos de optimización, se consideró kp igual a cero. Todos los demás parámetros afectan de manera apreciable al menos a una a las variables de estado del modelo, por lo tanto se mantuvieron.

Calibración de Parámetros

En primer lugar, se llevaron a cabo calibraciones con el conjunto total de 14 parámetros. La regresión resultó inestable en el sentido de que estimaciones muy diferentes entregaban resultados igualmente buenos. Esto se debe al efecto de los parámetros no identificables. Debido a lo anterior, se redujo el espacio paramétrico a nueve parámetros, lo que trajo consigo resultados estables. La Figura 2 muestra el desempeño de nuestra calibración con la curva resultante empleando OCFE para una de las condiciones de cultivo. De hecho, en todas las condiciones de cultivo se obtuvieron buenos ajustes a los datos experimentales. En la Tabla 1 se comparan los valores de la función objetivo usando el set de parámetros encontrados en este trabajo

En modelos biotecnológicos complejos es conveniente utilizar métodos de optimización global (GO) para encontrar el set de parámetros óptimos.

Condición	$fobj_{SMM}$	$fobj_{OCFE}$	$fobj_{GELMI}$
1	0,00638	0,00643	0,00876
2	0,01346	0,02405	0,01354
3	0,00489	0,00748	0,01570
4	0,00462	0,00553	0,00633

Tabla 1: Comparación del valor de la función objetivo para cada condición de cultivo

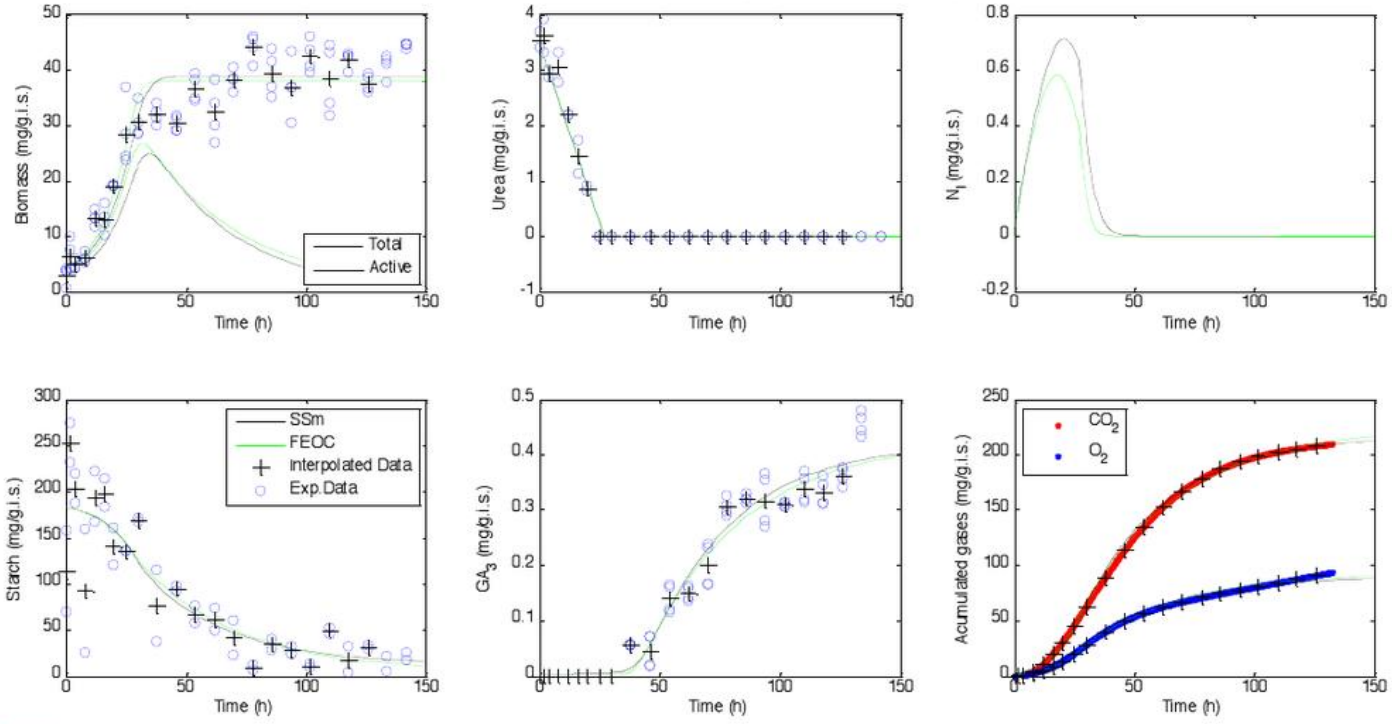


Figura 2: Calibración de parámetros para la condición de cultivo a 25°C y actividad de agua de 0,992

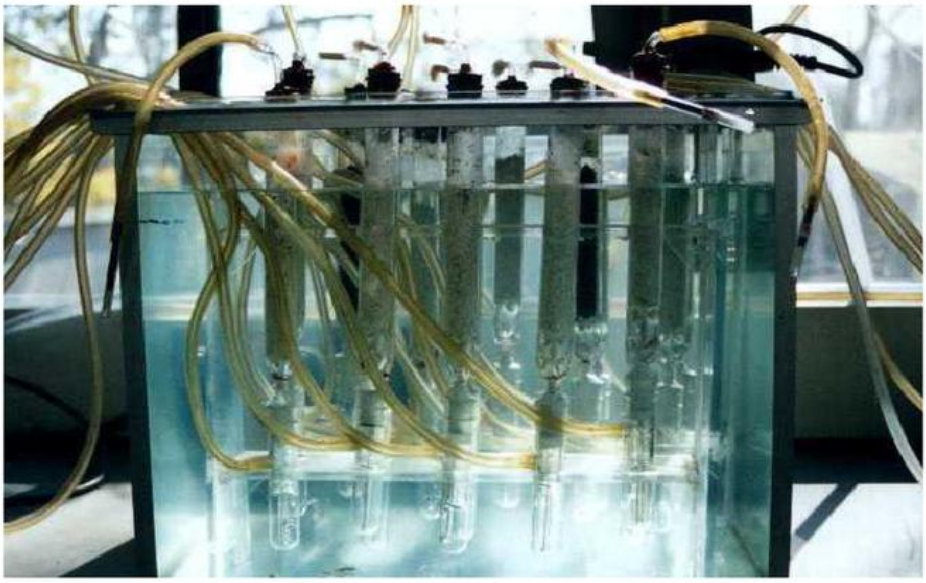
y otros reportados por Araya *et al.* [1] y Gelmi *et al.* [4].

Análisis de Significancia

El set reducido de parámetros generó mayores *t-values* e intervalos de confianza más pequeños para cada parámetro, lo que mejoró la confiabilidad del modelo. Sin embargo, los intervalos de confianza resultantes son sólo fiables si la confiabilidad de los valores fijados se ha demostrado en la literatura o por experimentos independientes. En general, la confiabilidad de los parámetros encontrados por Araya *et al.* [1] es comparable a la encontrada en esta investigación, y por ende, los parámetros tienen intervalos de confianza similares. Sólo se detectaron problemas de confiabilidad en el caso de la constante de inhibición para la producción de GA_3 (k_i) cuyo intervalo de confianza resultó ser muy grande en todas las condiciones de cultivo. Como solución, el análisis de sensibilidad nos puede ayudar a planificar experimentos futuros, de manera que este parámetro se pueda estimar de manera confiable. Una estrategia puede ser muestrear con mayor frecuencia en la zona donde dGA_3/dk_i es más sensible, es decir, $t > 50$ h.

CONCLUSIONES

El algoritmo de búsqueda dispersa demostró ser una herramienta eficaz y fácil para resolver el problema de calibración de un modelo biológico de crecimiento para el hongo filamentoso Gibberella fujikuroi en sustrato sólido. Los test estadísticos empleados resultaron ser cruciales para validar y modificar el modelo, ya que permitieron reducir el número de parámetros estimados, reduciendo la incertidumbre del modelo y mejorando la estabilidad de la estimación. Además, estos métodos de diagnóstico nos permitieron entregar recomendaciones con respecto a los tiempos de muestreo para experimentos adicionales, lo que ayudará a mejorar significativamente la confiabilidad y calidad del modelo para el desarrollo de mejores políticas de alimentación del cultivo para maximizar la producción de GA_3.



REFERENCIAS

1. ARAYA, Macarena M., ARRIETA, Juan J., PÉREZ-CORREA, Ricardo, BIEGLER, Lorenz T. and JORQUERA, Héctor. Fast and reliable calibration of solid substrate fermentation kinetic models using advanced non-linear programming techniques. Electronic Journal of Biotechnology. 10(1), 2007. DOI: 10.2225/vol10-issue5-fulltext-8
2. EGEA, José, RODRÍGUEZ-FERNÁNDEZ, María, BANGA, Julio R. and MARTÍ, Rafael. Scatter search for chemical and bio-process optimization. Journal of Global Optimization. 37(3):481-503, 2006.
3. GELMI, Claudio, PÉREZ-CORREA, Ricardo, GONZÁLEZ, M. and AGOSIN, Eduardo. Solid substrate cultivation of Gibberella fujikuroi on an inert support. Process Biochemistry. 35(10):1227-1233, 2000.
4. GELMI, Claudio, PÉREZ-CORREA, Ricardo and AGOSIN, Eduardo. Modelling Gibberella fujikuroi growth and GA3 production in solid-state fermentation. Process Biochemistry. 37(9):1033-1040, 2002.
5. SACHER, Johannes, SAA, Pedro, CÁRCAMO, Martín, LÓPEZ, Javiera, GELMI, Claudio and PÉREZ-CORREA, Ricardo. Improved calibration of a solid substrate model. Electronic Journal of Biotechnology. 14(5), 2011. DOI: 10.2225/vol14-issue5-fulltext-7

Principio Científico

Optimización: para realizar la calibración de los parámetros se minimizó la siguiente función de costos ponderada:

$$\min_{\Theta} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{X_{i,j}^{model} - X_{i,j}^{exp}}{n_i \cdot \bar{X}_i^{exp}} \right)^2$$

Donde Θ corresponde al espacio paramétrico, X^{exp} es la variable medida, X^{model} es la variable predicha, ni es el número de mediciones para la variable i y m es el número de variables.