CEA Comité Español de Automática

XLVI JORNADAS AUTOMÁTICA CARTAGENA 2025

UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Aprendizaje por refuerzo para el control del pH en fotobiorreactores de microalgas

Gil, Juan D.[a,*], Del Rio Chanona, Antonio[b], Guzmán, José L.[a], Berenguel, Manuel[a]

[a] *Departamento de Informática, CIESOL-ceiA3, Universidad de Almería, Ctra. Sacramento s/n, 04120, Almería, España.*
[b] *Sargent Centre for Process Systems Engineering, Imperial College London, SW7 2AZ, UK.*

To cite this article: Gil, Juan D., Del Rio Chanona, Antonio, Guzmán, José L., Berenguel, Manuel. 2025. Reinforcement Learning for pH Control in Microalgae Photobioreactors. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12099

Resumen

Este trabajo propone un sistema de control basado en aprendizaje por refuerzo para regular el pH en fotobiorreactores de microalgas, utilizando un agente basado en el algoritmo *Deep Deterministic Policy Gradient* (DDPG). Este enfoque aprende a partir de datos históricos generados por controladores convencionales, como el PID, sin necesidad de interacción directa con el sistema físico. Además, tras su implementación, el agente puede continuar entrenándose periódicamente con nuevas experiencias, lo que le permite adaptarse a las dinámicas cambiantes del proceso biológico. Los resultados en simulación muestran que el algoritmo propuesto mejora métricas de control tradicionales, como la integral del error absoluto en un 12 %, en comparación con un controlador PID. Asimismo, el reentrenamiento periódico favorece la adaptación y robustez del sistema. Estos resultados posicionan al aprendizaje por refuerzo como una alternativa prometedora para la automatización de este tipo de bioprocesos.

Palabras clave: Control adaptativo, Aprendizaje fuera de línea, Control de procesos biológicos.

Reinforcement Learning for pH Control in Microalgae Photobioreactors

Abstract

This work proposes a reinforcement learning control system to regulate the pH in microalgae photobioreactors, using an agent based on the *Deep Deterministic Policy Gradient* (DDPG) algorithm. This approach learns from historical data generated by conventional controllers, such as the PID, without requiring direct interaction with the real system. After its implementation, the agent can continue training periodically with new experiences, allowing it to adapt to the changing dynamics of the biological process. Simulation results show that the proposed algorithm improves traditional control metrics, such as the integral of absolute error, by 12 % compared to a PID controller. Additionally, periodic retraining supports the adaptation and robustness of the system. These results position reinforcement learning as a promising alternative for automating this type of bioprocess.

Keywords: Adaptive control, Offline learning, Biological process control.

1. Introducción

Las microalgas son organismos microscópicos capaces de realizar fotosíntesis y desarrollarse incluso en condiciones ambientales adversas. A través de la energía solar, transforman compuestos carbonados como el CO_2 en biomasa y generan una notable cantidad de oxígeno (Guzmán et al., 2021). Para favorecer su crecimiento, es necesario suministrar nutrientes adicionales como carbono, nitrógeno y fósforo. El carbono, usualmente proporcionado mediante la inyección de CO_2, también cumple la función de regular el pH del sistema, una de las variables más relevantes y delicadas de controlar. Por su parte, el nitrógeno y el fósforo pueden añadirse directamente al medio o ser asimilados desde el propio entorno, especialmente si se utilizan aguas residuales.

Centrándose en el pH, esta variable destaca como una de las más críticas, ya que influye directamente en la solubilidad y disponibilidad tanto del CO_2 como de los nutrientes,

*Autor para correspondencia: juandiego.gil@ual.es

afectando de manera considerable el metabolismo de las microalgas (Juneja et al., 2013). La propia fotosíntesis genera fluctuaciones constantes en el pH, lo que complica aún más su control. Generalmente, se utilizan sistemas de control tipo todo/nada, los cuales no consideran la dinámica del sistema ni las perturbaciones. El uso de este tipo de estrategias de control se debe, en gran parte, a la dificultad para desarrollar un modelo que represente con precisión la dinámica del proceso (Guzmán et al., 2021).

Dada la dificultad de modelado en este tipo de procesos, las técnicas de control adaptativo han emergido como una alternativa prometedora en la literatura. Diversos estudios han abordado esta problemática desde distintos enfoques. En el trabajo de Caparroz et al. (2024), se presentó un modelo adaptativo basado en un árbol de regresión capaz de predecir el pH en diferentes condiciones operacionales. Posteriormente, este árbol se usó para el desarrollo e implementación de una estrategia de control adaptativa basada en un controlador Proporcional, Integral, y Derivativo (PID). Por otro lado, en el trabajo de Caparroz et al. (2023) se exploró el uso de una estrategia adaptativa por modelo de referencia (MRAC, por sus siglas en inglés *Model Reference Adaptive Control*) para el control del pH. Esta estrategia se hibridó posteriormente con un controlador de tipo PID (Caparroz et al., 2025). No obstante, este tipo de estrategias adaptativas aún presentan algunos desafíos relacionados con la necesidad de modelos nominales, y su limitada capacidad de generalización frente a dinámicas altamente no lineales y cambiantes, como las que se presentan en un fotobiorreactor durante su operación anual.

Ante las limitaciones de los enfoques adaptativos, el uso de técnicas de aprendizaje basadas en datos se presenta como una alternativa más flexible. Uno de los primeros trabajos en este sentido fue el presentado por Pataro et al. (2023). En este estudió se desarrolló una estrategia de control predictivo basada en modelo (MPC, por sus siglas en inglés *Model Predictive Control*) combinada con una función oráculo que aprende directamente de datos en línea para ajustar las incertidumbres del modelo nominal usado intrínsecamente en el MPC. La estrategia proporcionó buenos resultados en términos de adaptación y control a las diferentes dinámicas inducidas por diferentes tipos de medio de cultivo. No obstante, a pesar de los buenos resultados obtenidos con este enfoque, esta estrategia sigue dependiendo de estructuras de control explícitas y de cierto conocimiento previo sobre la dinámica del sistema. En este contexto, el uso de técnicas de aprendizaje por Refuerzo (RL, por sus siglas en inglés *Reinforcement Learning*) representa un paso adicional hacia una mayor autonomía y adaptabilidad del controlador. A diferencia del MPC, que optimiza decisiones en función de un modelo predictivo, el RL aprende directamente una política de control a partir de la interacción con el entorno o de experiencias previas recogidas en datos históricos, lo que le permite ajustar su comportamiento de forma continua sin requerir una representación explícita del modelo del sistema (Petsagkourakis et al., 2020).

De una forma generalizada, las estrategias de RL pueden clasificarse como *on-policy* u *off-policy*, según si la política utilizada para generar los datos de entrenamiento coincide o no con la política que se está evaluando o mejorando. En los enfoques *on-policy*, el agente de RL aprende a partir de las acciones que toma siguiendo su propia política (Sachio et al., 2021). Por su parte, el enfoque *off-policy* permite al agente aprender sobre una política distinta a la que generó con sus propias acciones (Monteiro and Kontoravdi, 2024), lo que ofrece mayor flexibilidad y reutilización de experiencias pasadas. Esta última característica resulta especialmente valiosa en el caso de los fotobiorreactores, donde las pruebas en línea pueden ser costosas, lentas o incluso inviables debido a la naturaleza biológica y a la sensibilidad del proceso. Así, los métodos *off-policy* permiten aprovechar datos históricos generados bajo diferentes políticas de control en bucle cerrado, lo que reduce la necesidad de realizar experimentación directa y acelera el proceso de entrenamiento del agente.

En este trabajo se propone un sistema de control basado en aprendizaje por refuerzo *off-policy* para la regulación del pH en fotobiorreactores de microalgas. En concreto, se propone un algoritmo de control basado en un agente *Deep Deterministic Policy Gradient* (DDPG) (Castilla et al., 2025; Rajasekhar et al., 2025). Esta estrategia aprende a partir de experiencias históricas generadas mediante controladores convencionales, como PIDs, sin requerir interacción directa con el sistema real. Una vez implementado, el agente puede continuar su entrenamiento de forma periódica con nuevas experiencias siguiendo las ideas presentadas por Wang et al. (2025), lo que le permite adaptarse a las dinámicas variables propias de los procesos biológicos. Las ventajas del enfoque propuesto se validan mediante un estudio comparativo en simulación, utilizando un modelo de primeros principios alimentado con datos reales de las instalaciones del convenio UAL-IFAPA. Además, se emplean datos recolectados en distintas épocas del año para entrenar y evaluar el desempeño del agente bajo condiciones diversas, lo que demuestra sus capacidades de adaptación.

2. Materiales y métodos

2.1. Descripción del sistema

El fotobiorreactor *raceway* utilizado como referencia en este estudio tiene una superficie de 80 m^2 y está ubicado en el centro IFAPA de la Junta de Andalucía, cerca de la Universidad de Almería (ver Fig. 1). Este reactor consta de dos canales de 40 metros de largo, 1 metro de ancho y 0.3 metros de profundidad. El proceso de mezcla y circulación se lleva a cabo mediante un sistema de palas de 1.2 metros de diámetro, equipada con 8 palas. Después de la rueda, se encuentra un foso donde se inyecta CO_2 y aire, utilizados para controlar el pH y el oxígeno disuelto, respectivamente.

Figura 1: Reactores *raceway* disponibles en las instalaciones del Convenio UAL-IFAPA.

El sistema cuenta con una instrumentación completa que permite la recolección de datos en tiempo real, registrando cada segundo diversas variables del proceso como el pH,

oxígeno disuelto, temperatura del cultivo y nivel, así como parámetros ambientales como radiación solar, temperatura del aire, velocidad del viento y humedad relativa, entre otros. Las mediciones de pH y oxígeno disuelto se realizan en dos puntos clave: justo después del foso y al final del canal, justo antes de las palas. Este último punto representa el mayor desafío en términos de control, ya que se encuentra más alejado del área de inyección de CO_2 y, por tanto, es el foco principal de las estrategias de regulación implementadas en el sistema. En Caparroz et al. (2025) se puede encontrar una descripción más completa y detallada del sistema.

2.2. Modelo del sistema

2.2.1. Modelo biólogico

En esta sección del modelo se aborda la velocidad de fotosíntesis, la cual define el crecimiento de las microalgas y se asocia con la tasa de producción de oxígeno por unidad de biomasa. Para ello, se emplea el cálculo de la tasa de crecimiento específica (μ). Dicha tasa se ve influida por múltiples variables, entre ellas: la disponibilidad de luz (I_{av}), la temperatura (T), el pH (pH), la concentración de oxígeno disuelto (O_2) y la respiración microbiana (m). El término relacionado con la luz, $\mu(I_{av})$, se ajusta mediante tres factores ($\overline{\mu(\cdot)}$) que toman valores entre 0 y 1 y dependen de la temperatura, el pH y la concentración de oxígeno. También se considera que los nutrientes como nitrógeno, fósforo y otros están en exceso, por lo que no se incluyen explícitamente. La expresión que describe la tasa de crecimiento específico se puede expresar como:

$$\mu = \mu(I_{av}) \cdot \overline{\mu(T)} \cdot \overline{\mu(pH)} \cdot \overline{\mu(O_2)} - m. \tag{1}$$

El desarrollo de cada unos de los términos de la ecuación no se ha incluido en este trabajo debido a la falta de espacio. No obstante, el lector interesado puede consultar una descripción completa de las ecuaciones y términos que modelan estos parámetros en (Guzmán et al., 2021; Caparroz et al., 2023).

2.2.2. Modelo dinámico

Esta parte del modelo incluye algunos de los balances de masa y energía dentro de las diferentes zonas del fotobiorreactor, con el fin de evaluar cómo influyen distintas variables en la tasa de crecimiento del cultivo. Se asume que el foso actúa como un sistema de mezcla perfecta, por lo que su comportamiento se describe mediante ecuaciones diferenciales ordinarias. Las principales variables del modelo son la concentración de biomasa (C_b), la concentración de oxígeno disuelto ($[O_2]$) y la concentración de carbono inorgánico total ($[C_T]$), cuyos balances de masa se muestran a continuación:

$$\frac{dC_b(t)}{dt} = C_b(t) \cdot \left(\mu - d_r - \frac{Q_d}{V_r}\right), \tag{2}$$

$$\begin{aligned}\frac{d[O_2](t)}{dt} = {} & \frac{Q_d}{V_r} \cdot ([O_2^*] - [O_2](t)) + \mu \cdot \frac{Y_{O_2}}{M_{O_2}} \cdot C_b(t) \\ & + K_{laO_2atm} \cdot ([O_2^*] - [O_2](t)) \\ & + K_{laO_2} \cdot \left([O_2^{iny}] - [O_2](t)\right),\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{d[C_T](t)}{dt} = {} & \frac{Q_d}{V_r} \cdot ([C_{Tin}] - [C_T](t)) - \mu \cdot \frac{Y_{CO_2}}{M_{CO_2}} \cdot C_b(t) \\ & + K_{laCO_2atm} \cdot ([CO_2^*] - [CO_2](t)) \\ & + K_{laCO_2} \cdot \left([CO_2^{iny}] - [CO_2](t)\right).\end{aligned} \tag{4}$$

La concentración de biomasa (C_b) en el reactor está gobernada por la tasa de crecimiento microbiano (μ), la tasa de mortalidad (d_r) y la relación entre el flujo de dilución (Q_d) y el volumen del reactor (V_r), como se describe en la ecuación (2). La dinámica del oxígeno disuelto se ve influenciada por cuatro componentes principales (ver ecuación (3)): i) efecto de la dilución, considerando la concentración de equilibrio atmosférico, ii) producción de O_2 por las microalgas, donde Y_{O_2} representa el rendimiento de oxígeno por biomasa y M_{O_2} su masa molar, iii) transferencia de masa entre el reactor y la atmósfera, y iv) intercambio gaseoso con el aire inyectado (asumiendo $[O_2^{inv}] = [O_2^*]$ para aire ambiental). Para el caso del carbono inorgánico total, la ecuación es análoga a la anterior pero con signo opuesto en el término fotosintético (ver ecuación (4)), reflejando el consumo de CO_2 (a diferencia de la producción de O_2). Nótese que la transferencia de masa de ambos gases depende de sus coeficientes volumétricos ($K_{la_{O_2}}$, $K_{la_{CO_2}}$) y del gradiente de concentración respecto al equilibrio.

2.3. Sistema de control por RL

2.3.1. Conceptos previos

La mayoría de los algoritmos de RL suponen que el entorno puede modelarse como un proceso de decisión de Markov (MDP, por sus siglas en inglés *Markov Decision Process*), representado formalmente como una quíntupla $\mathcal{S}, \mathcal{A}, P, R, \gamma$, donde $\mathcal{S}$ es el conjunto de estados del sistema, $\mathcal{A}$ el conjunto de acciones disponibles, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ la función de probabilidad de transición entre estados, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ la función de recompensa y γ el factor de descuento. No obstante, en entornos reales, el estado completo del sistema no suele estar disponible, por lo que se trabaja con observaciones (denotadas por O) que ofrecen una visión parcial del estado real, dando lugar a un entorno parcialmente observable MDP (POMDP). Un agente de RL busca aprender una política óptima basada en las observaciones disponibles, siguiendo el esquema de comunicación con el sistema real (o entorno) que se muestra en la Fig. 2.

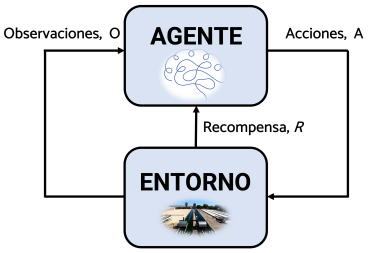

Figura 2: Flujo de trabajo simplifacdo de un sistema de RL.

Una arquitectura ampliamente utilizada en este contexto es la de *actor-crítico*, que combina un *actor*, responsable de seleccionar acciones a_i ante una observación o_i siguiendo una política determinista definida por la función $\pi(o_i; \boldsymbol{\theta})$, y un *crítico*, que evalúa dichas acciones a través de una función de valor de acción $Q(o_i, a_i; \boldsymbol{\Phi})$. Nótese que en la formulación previa, los parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Phi}$ representan los pesos de los modelos que definen el comportamiento del actor y del crítico, respectivamente, y son ajustados durante el entrenamiento para optimizar tanto la política de decisión como su evaluación.

2.3.2. *Deep Deterministic Policy Gradient*

Entre los algoritmos más representativos que adoptan la estructura *actor-crítico* en entornos con espacios de acción continuos se encuentra el DDPG (Rajasekhar et al., 2025). Este algoritmo utiliza dos redes neuronales: una red *actor* que genera acciones deterministas y una red *crítico* que estima el valor de acción.

Para el entrenamiento fuera de línea de un agente siguiendo este algoritmo, se almacenan experiencias previas en una memoria conocida como *buffer* de experiencias, la cual está compuesta por medidas de las observaciones, acciones y recompensas para determinados tiempos de muestreo. Posteriormente, durante el entrenamiento, se extraen mini-lotes aleatorios de tamaño M del *buffer* para actualizar los modelos del *actor* y el *crítico*. En concreto, el *crítico* se entrena minimizando la siguiente función de error:

$$y_i = r_i + \gamma Q_t(o_i, \pi_t(o_i; \boldsymbol{\theta}_t); \boldsymbol{\Phi}_t), \tag{5}$$

$$L(\boldsymbol{\Phi}) = \frac{1}{M} \sum_{i=1}^{M} (y_i - Q_t(o_i, a_i; \boldsymbol{\Phi}))^2 . \tag{6}$$

El valor objetivo y_i se calcula sumando la recompensa inmediata r_i y la recompensa futura descontada, estimada mediante redes objetivo desacopladas (conocidas como *target networks*). Estas redes, denotadas como Q_t y π_t, son copias lentas de las redes principales que se actualizan de forma suave para estabilizar el aprendizaje y evitar oscilaciones en las estimaciones del valor (Lillicrap et al., 2015).

El *actor*, por su parte, se actualiza maximizando la recompensa acumulada esperada a través del siguiente gradiente estimado sobre un mini-lote de tamaño M:

$$\nabla_{\boldsymbol{\theta}} J \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{G}_{a_i} \mathbf{G}_{\pi_i}, \tag{7}$$

donde $\mathbf{G}_{a_i}$ representa el gradiente de la salida del *crítico* con respecto a la acción generada por la red *actor*, y $\mathbf{G}_{\pi_i}$ corresponde al gradiente de la salida del *actor* con respecto a sus propios parámetros.

2.3.3. *Algoritmo de RL propuesto*

En este trabajo, se desarrolla un marco de aprendizaje por refuerzo fuera de línea aplicando el algoritmo DDPG, con el objetivo de desarrollar una estrategia de control para la variable de pH en fotobiorreactores de microalgas. La metodología consta de tres pasos detallados en el Algoritmo 1.

Como se puede apreciar, el proceso comienza con la preparación de un conjunto de datos históricos del bioproceso, que incluye observaciones, acciones y recompensas recolectadas bajo diversas condiciones de operación. Este conjunto se utiliza para entrenar fuera de línea una política DDPG, que luego se ajusta finamente en línea para adaptarse al entorno en tiempo real siguiendo las ideas propuestas en Wang et al. (2025). Así, la política combina el conocimiento aprendido con datos históricos y la adaptación continua al sistema real.

3. Resultados

3.1. *Recolección de experiencias históricas*

Una ventaja clave de los algoritmos de aprendizaje por refuerzo *off-policy* es su capacidad para aprender a partir de experiencias previas, como las generadas por controladores PID, comunes en fotobiorreactores por su robustez (Guzmán et al., 2021). Por ello, el agente DDPG se entrenó con datos obtenidos operando el sistema con un controlador PID, según el esquema de la Fig. 3.

Algoritmo 1: Desarrollo del agente DDPG en un entorno fuera de línea con ajuste fino en tiempo real

Entrada: Conjunto de datos históricos $\mathcal{D}$ del bioproceso con diversas condiciones operativas.
Salida: Política de control robusta y optimizada para la regulación del pH en el biorreactor.
Paso 1: Preparación del conjunto de datos históricos $\mathcal{D}$:

1. Recolectar los datos históricos de operación del sistema.
2. Preprocesar los datos recopilados y organizarlos en tuplas de observaciones-acciones-recompensas.

Paso 2: Entrenamiento fuera de línea del agente DDPG:

1. Inicializar aleatoriamente los parámetros de las redes neuronales del agente DDPG.
2. Aplicar el proceso de entrenamiento detallado en la sección 2.3.2.

Paso 3: Ajuste fino en línea del agente DDPG:

1. Inicializar el *buffer* de experiencias utilizando el conjunto de datos históricos $\mathcal{D}$.
2. Recolectar nuevas experiencias en línea de forma continua durante la operación del bioproceso.
3. Actualizar dinámicamente el *buffer* de experiencias, manteniendo las experiencias más recientes.
4. Realizar el ajuste fino del agente DDPG mediante el proceso de entrenamiento descrito en la sección 2.3.2.

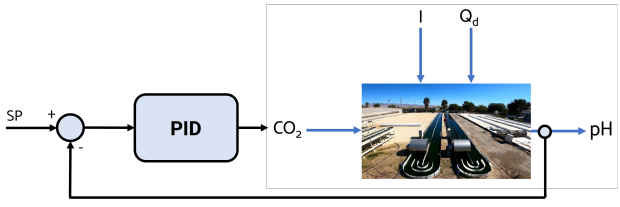

Figura 3: Esquema de control PID para los fotobiorreactores.

Como se puede apreciar, en el esquema de control se considera el caudal de CO_2 como la variable manipulable para el control del pH. Además, también se representan las dos principales variables que influencian el comportamiento dinámico del sistema, las cuales son la radiación (I) y el caudal de dilución (Q_d). Nótese que el controlador PID ha sido configurado usando la forma ideal sin la parte derivativa, con los parámetros $K_p = -55$ [m^3/s], y K_i =-0.05 [s^{-1}], y tiempo de muestreo 1 segundo. Los datos reales usados para la obtención de estas experiencias corresponden a los días 17, 18, y 19 de abril del año 2023. Los resultados se presentan en la Fig. 4.

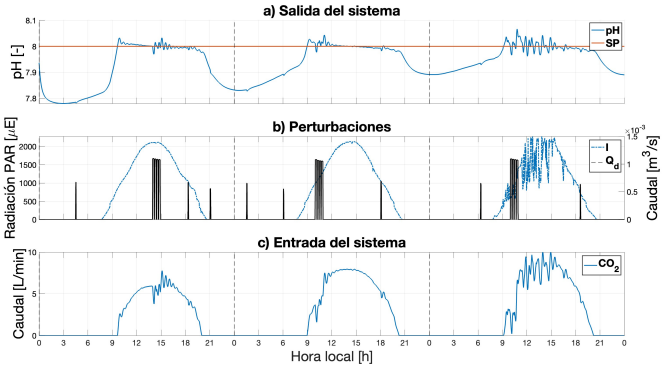

Figura 4: Resultados obtenidos con el controlador PID.

3.2. *Configuración del controlador RL y entrenamiento fuera de línea*

Una vez obtenidos los datos históricos para el entrenamiento fuera de línea del agente, se procede a la configuración de este y a su entrenamiento. Al usar el agente para labores de control, se considera la siguiente función de recompensa:

$$r_i = -(pH_i - \mathrm{SP})^2, \tag{8}$$

donde SP representa la referencia de pH que se desea seguir, con un valor de 8 para el tipo de cepa utilizado. Además, en cuanto al conjunto de observaciones, este se compone de medidas de las variables de pH, radiación (I), caudal de dilución ($\mathbf{Q_d}$), caudal de CO_2, error entre la referencia (SP) y el pH y la integral de este error.

Nótese que la inclusión de las variables de radiación (I) y caudal de dilución ($\mathbf{Q_d}$) en el conjunto de observaciones permite obtener una acción de control por adelanto de forma intrínseca en el agente de RL. Asimismo, el uso del error y de su integral aporta al agente una noción explícita del desvío frente al objetivo y de su acumulación a lo largo del tiempo, dos elementos fundamentales en el diseño de estrategias de control para sistemas dinámicos. En este caso, estas variables adquieren una relevancia aún mayor, ya que el entrenamiento se realiza a partir de datos generados por un controlador PID en bucle cerrado, cuya ley de control se fundamenta precisamente en el error y su integral. De este modo, se facilita que el agente de RL aprenda una política de control eficaz a partir de la lógica implícita en el comportamiento del sistema experto del que ha aprendido, es decir, del sistema PID. El esquema general del controlador RL resultante se muestra en la Fig. 5.

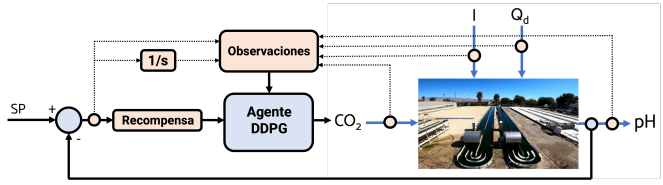

Figura 5: Esquema de control RL para los fotobiorreactores.

Se debe hacer notar que la implementación del agente se ha desarrollado en MATLAB (The MathWorks, Inc., 2024). La red neuronal del actor se ha configurado con 8 capas con diferentes características como *ReLU* o *Tanh*. Por su parte, el crítico se ha configurado con 9 capas también de diferentes características. Para el entrenamiento fuera de línea del agente, se ha considerado un tiempo de muestreo de 1 segundo para el cálculo de la función de recompensa y acumulación de observaciones, imitando el muestreo usado en la estrategia PID. Además, el entorno de entrenamiento se configuró con un total de hasta 1000 épocas (pasadas completas sobre el conjunto de datos para la optimización de parámetros de las redes neuronales del agente).

3.3. *Implementación del agente y estudio comparativo*

Una vez entrenado el agente con los datos de experiencias previas, se implementó para el control del pH en el simulador. Con el fin de evaluar su desempeño en condiciones no cubiertas durante el entrenamiento, se utilizaron datos de una época del año distinta, lo que permitió que la dinámica y las condiciones operativas del sistema fueran considerablemente diferentes a las empleadas en el entrenamiento fuera de línea. Específicamente, los datos provienen de los días 25 y 26 de noviembre de 2023. Además, se plantea un estudio comparativo utilizando tres estrategias: i) una estrategia de control basada en un PID nominal (denotado como PID), configurado como se detalla en la sección 3.1, con el objetivo de comparar el rendimiento del agente con el del sistema experto utilizado durante su entrenamiento, ii) una estrategia de control utilizando el agente DDPG entrenado fuera de línea (denotado como DDPG), sin ajuste fino en línea, y iii) una estrategia que incorpora un ajuste fino de los parámetros del agente al final del día (referido como DDPG-AF), integrando los datos históricos en el *buffer* de experiencias con los adquiridos a lo largo de la operación diaria. Los resultados se muestran en la Fig. 6 y de forma cuantitativa en la tabla 1. Las métricas que se presentan en la tabla se calculan de la siguiente forma:

$$\mathrm{IAE} = \int_0^T |e(t)|\, dt, \quad \mathrm{EC} = \int_0^T |\Delta u(t)|\, dt. \tag{9}$$

Como se puede apreciar en la gráfica, las condiciones en las que se plantea el estudio comparativo son más desafiantes que las usadas para el entrenamiento debido a las múltiples perturbaciones en el caudal de dilución (ver Fig. 6-b). Bajo estas condiciones, y centrándonos en el primer día de operación, las actuaciones de las tres estrategias son bastante similares. De hecho, al no haber reajuste fino hasta el final del día, la actuación del agente DDPG y DDPG-AF es igual. Los resultados de este primer día demuestran que, aunque se ha entrenado al agente con datos de otra época del año y otras condiciones de operación, este ha aprendido una política que captura patrones fundamentales del entorno y le permite desenvolverse adecuadamente en condiciones distintas a las del entrenamiento.

Las principales diferencias se presentan en el segundo día. Tras realizar el ajuste fino en el agente DDPG con los datos del día previo, se puede observar cómo el controlador DDPG-AF mejora notablemente los resultados tanto del PID, como del agente DDPG entrenado fuera de línea. Aparte de los resultados gráficos mostrados en la Fig. 6, la comparación cuantitativa de la tabla 1 muestra cómo el DDPG-AF mejora los resultados en términos de IAE entorno a un 12 % respecto a la estrategia PID, y entorno a un 5 % respecto al DDPG. Además, también se mejoran los resultados en términos de esfuerzo de control, debido a su acción de control por adelanto intrínseca.

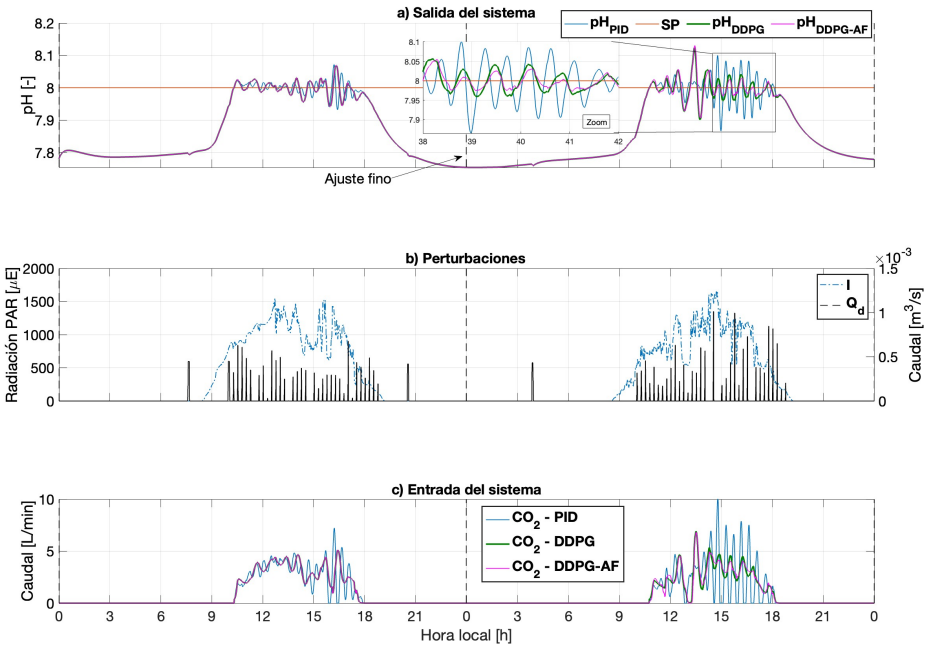

Figura 6: Comparativa de controladores.

Tabla 1: Comparación de métricas de control

Controlador	IAE	EC
PID	68.5	222.9
DDPG	63.7	96.5
DDPG-AF	60.8	91.8

4. Conclusiones

Este trabajo presenta un sistema de control basado en aprendizaje por refuerzo para la regulación del pH en fotobiorreactores de microalgas. El enfoque se fundamenta en un agente DDPG entrenado inicialmente con datos históricos generados por un sistema experto, lo que permite aprender una política de control efectiva sin requerir interacción directa con el entorno. Posteriormente, el agente continúa refinando su comportamiento mediante la experiencia obtenida durante la operación, lo que le confiere capacidad de adaptación frente a las variaciones del proceso biológico. Los resultados obtenidos muestran que se pueden mejorar métricas clásicas de control como el IAE en torno a un 12 % en comparación con un controlador PID. Además, el agente trata de forma intrínseca perturbaciones, lo que permite mejorar también los resultados en términos de esfuerzo de control. En general, estos resultados posicionan al RL como una alternativa prometedora para la automatización de este tipo de bioprocesos.

Agradecimientos

Esta publicación es parte del proyecto de I+D+i PID2023-150739OB-I00, financiado/a por MCIN/ AEI/10.13039/501100011033/ y "FEDER Una manera de hacer Europa". Además, cuenta con financiación proveniente del Ministerio de Ciencia, Innovación y Universidades a través del programa de estancias de movilidad en centros extranjeros de enseñanza superior e investigación.

Referencias

Caparroz, M., Guzmán, J. L., Berenguel, M., Acién, F., 2024. A novel data-driven model for prediction and adaptive control of pH in raceway reactor for microalgae cultivation. New Biotechnology 82, 1–13.

Caparroz, M., Guzmán, J. L., Berenguel, M., Gil, J. D., Acién, F. G., 2023. Model reference adaptive control for pH regulation. Revista Iberoamericana de Automática e Informática industrial 22 (2), 126–134.

Caparroz, M., Guzmán, J. L., Gil, J. D., Berenguel, M., Acién, F. G., 2025. A hybrid MRAC-PI approach to regulate pH in raceway reactors for microalgae production. Control Engineering Practice 156, 106191.

Castilla, M. M., Campoy-Iniesta, C., Álvarez, J. D., 2025. Control del confort térmico mediante aprendizaje por refuerzo en edificios. Revista Iberoamericana de Automática e Informática Industrial (RIAI) 22 (2), 146–155.

Guzmán, J. L., Acién, F., Berenguel, M., 2021. Modelling and control of microalgae production in industrial photobioreactors. Revista Iberoamericana de Automática e Informática Industrial 18 (1), 1–18.

Juneja, A., Ceballos, R. M., Murthy, G. S., 2013. Effects of environmental factors and nutrient availability on the biochemical composition of algae for biofuels production: a review. Energies 6 (9), 4607–4638.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Monteiro, M., Kontoravdi, C., 2024. Bioprocess control: A shift in methodology towards reinforcement learning. In: Computer Aided Chemical Engineering. Vol. 53. Elsevier, pp. 2851–2856.

Pataro, I. M., Gil, J. D., Guzmán, J. L., Berenguel, M., Lemos, J. M., 2023. A learning-based model predictive strategy for pH control in raceway photobioreactors with freshwater and wastewater cultivation media. Control Engineering Practice 138, 105619.

Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D., del Rio-Chanona, E. A., 2020. Reinforcement learning for batch bioprocess optimization. Computers & Chemical Engineering 133, 106649.

Rajasekhar, N., Radhakrishnan, T., Samsudeen, N., 2025. Exploring reinforcement learning in process control: a comprehensive survey. International Journal of Systems Science, 1–30.

Sachio, S., del Rio-Chanona, E. A., Petsagkourakis, P., 2021. Simultaneous process design and control optimization using reinforcement learning. IFAC-PapersOnLine 54 (3), 510–515.

The MathWorks, Inc., 2024. MATLAB R2024b. `http://es.mathworks.com/products/matlab/`, accessed on 15/05/2025.

Wang, H., Kontoravdi, C., Del Rio Chanona, A., 2025. Offline reinforcement learning for bioprocess optimization with historical data. In: 14th IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems (DYCOPS 2025): Slovakia, Bratislava, June 16-19, 2025.